DOCUMENT RESUME

ED 294 890 TM 011 469

AUTHOR Natriello, Gary
TITLE Evaluation Processes in Schools and Classrooms. Report No. 12. May, 1987.
INSTITUTION Johns Hopkins Univ., Baltimore, Md. Center for Social Organization of Schools.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE May 87
GRANT OERI-G-86-0006
NOTE 98p.
PUB TYPE Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Academic Achievement; *Classroom Research; Elementary Secondary Education; Evaluation Criteria; *Evaluation Methods; Literature Reviews; *Models; Research Design; Student Evaluation

ABSTRACT

Literature relating to evaluation processes in schools and classrooms is reviewed to develop a conceptual framework for integrating research on such evaluation processes. This report was prepared by the Middle School program. Commentary and research on elements of the evaluation process are examined in this framework, and the ways in which formal programs and policies have impact on evaluation are considered. The framework here presented is summarized as: (1) establishing evaluation purposes; (2) assigning tasks to students; (3) setting student performance criteria; (4) setting performance standards; (5) sampling information on student performance; (6) assessing student performance; (7) providing feedback to students; and (8) monitoring outcomes of evaluation. This framework is a first step toward the comprehensive framework needed for effective evaluation. A 13-page list of references and four tables are presented. (SLD)

Center for Research On Elementary & Middle Schools

Report No. 12

May, 1987

EVALUATION PROCESSES IN SCHOOLS AND CLASSROOMS

Gary Natriello

Evaluation Processes in Schools and Classrooms

Grant No. OERI-G-86-0006

Gary Natriello

Teachers College, Columbia University

Report No. 12

May 1987

Published by the Center for Research on Elementary and Middle Schools, supported as a national research and development center by funds from the Office of Educational Research and Improvement, U.S. Department of Education. The opinions expressed in this publication do not necessarily reflect the position or policy of the OERI, and no official endorsement should be inferred.

Center for Research on Elementary and Middle Schools
The Johns Hopkins University
3505 North Charles Street
Baltimore, Maryland 21218

Printed and assembled by:
VSP Industries
2440 West Belvedere Avenue
Baltimore, Maryland 21215

The Center

The mission of the Center for Research on Elementary and Middle Schools is to produce useful knowledge about how elementary and middle schools can foster growth in students' learning and development, to develop and evaluate practical methods for improving the effectiveness of elementary and middle schools based on existing and new research findings, and to develop and evaluate specific strategies to help schools implement effective research-based school and classroom practices.

The Center conducts its research in three program areas: (1) Elementary Schools, (2) Middle Schools, and (3) School Improvement.

The Elementary School Program

This program works from a strong existing research base to develop, evaluate, and disseminate effective elementary school and classroom practices; synthesizes current knowledge; and analyzes survey and descriptive data to expand the knowledge base in effective elementary education.

The Middle School Program

This program's research links current knowledge about early adolescence as a stage of human development to school organization and classroom policies and practices for effective middle schools. The major task is to establish a research base to identify specific problem areas and promising practices in middle schools that will contribute to effective policy decisions and the development of effective school and classroom practices.

School Improvement Program

This program focuses on improving the organizational performance of schools in adopting and adapting innovations and developing school capacity for change.

This report, prepared by the Middle School program, develops a model for understanding and improving student evaluation processes in schools and classrooms, and reviews research on evaluation processes in terms of the model.

Abstract

This paper reviews literature relating to evaluation processes in schools and classrooms. The review provides a conceptual framework to integrate research on evaluation processes in schools and classrooms, examines commentary and research on elements of the evaluation process, and seeks to provide an understanding of how formal programs and policies affect evaluation processes.

Introduction

The evaluation of student performance is a central task of schools and teachers. Indeed, evaluation activities permeate the educational process. Although this is now particularly apparent, as schools are under increased pressure for greater accountability and improved performance, the pressure on and interest in evaluation processes is nothing new. Throughout the history of American education, evaluation of student performance has been an element of enduring concern to educators, to students, and to parents (Crooks, 1933), and social scientists and educators have amassed a considerable body of research and commentary related to the evaluation process.

Such work appears under a number of different rubrics -- from testing, accountability, and standards to incentives, grading, and marking. Evaluation processes include those initiated and directed by teachers as well as those sponsored by the school, the school district, accrediting agencies, and state and federal governments.

This review 1) provides a conceptual framework to integrate research on evaluation processes in schools and classrooms; 2) examines commentary and research on elements of the evaluation process in terms of that conceptual framework; and 3) develops some understanding of the ways in which formal programs and policies have an impact on evaluation processes in schools and classrooms.

Conceptual Framework for School and Classroom Evaluation Processes

Evaluation processes can be conceptualized in many ways. For example, evaluation might be considered as an interpersonal process with important implications for individual motivation, as a social and organizational process with substantial effects on social and institutional stability, or as a political process with an impact on the distribution of power and resources in a system (Natriello, 1985). A framework adopted to consider evaluation processes in schools and classrooms might contain elements of each of these approaches. We will emphasize a framework which permits the organization and presentation of theory and findings on how educators structure the process of evaluation and the likely outcomes of the structure that is adopted.

Figure 1 depicts the key elements in the framework for considering evaluation processes in schools and classrooms.

Insert Figure 1 About Here

The purposes of student evaluation can be many and varied, and play an important role in determining the nature of the evaluation activities. The assignment of academic tasks to students sets the stage for the evaluation activities that follow. Through the process of assignment, students are put on notice that they are

expected to perform a certain task. But to attempt to respond to teachers' expectations, they need information on the nature of the desired performance -- they need criteria that are specified for the task performance which tell them what aspects of the performance are important to the teacher. Information for task performance also comes from standards that communicate the level of performance that students are supposed to achieve. With tasks assigned and criteria and standards established and communicated, students are in a position to engage in the appropriate activities.

Collecting information on student performance of assigned tasks and the outcomes of those tasks involves a sampling process, because total information is typically impractical or impossible to collect. The sample of information on student performance may be used in conjunction with the criteria and standards as evaluators actually develop an appraisal of the student performance. Once the appraisal is developed, it still remains for the evaluator to communicate the results of the appraisal to the student performer. The feedback process might then lead to a number of outcomes which, presumably, relate to the original purposes of the evaluation process.

This model of the evaluation process lays out in a generic way the various elements of evaluation. It is not unlike other models of evaluation and control processes (e.g., Lawler, 1976) and models of cybernetic feedback processes (e.g., Bloom, 1980). But it also suggests how the various elements may be related to one another. For example, the purposes of the evaluation process are likely to

influence how tasks are assigned, the kinds of criteria that are set, how samples of student performance are collected, the appraisal process, and the nature of the feedback provided to students. But the model does not suggest that the stages must take place only in the order portrayed. The circular arrangement of elements conveys the notion that certain evaluation procedures are adopted for historical and idiosyncratic reasons that may have little to do with other procedures. For example, the sampling of student performance might derive from established procedures that limit the purposes to which the evaluation can be put. (Glaser, 1963, observes the inappropriate use of norm-referenced tests to assess the effects of educational programs.) So too, the mechanisms for providing feedback to students may stem from tradition and provide poor information for assessing performance in terms of certain criteria. Critics of traditional report cards have charged that they provide little insight for students and parents interested in working to improve performance (Giannangelo and Lee, 1974). Thus, the model of elements of the evaluation process describes a rational progression for the process, but also reveals the somewhat less than rational nature of the process as it operates in schools and classrooms.

The Purposes of Student Evaluation in Schools and Classrooms

Aside from the obligatory brief section on the purposes of evaluation at the front of texts on measurement and evaluation (e.g., Remmers, Gage, and Rumme, 1960; Lien, 1967; Ahmann and Glock, 1967), the purposes of evaluation receive scant attention, which is parti-

cularly ironic in a literature that encourages teachers to specify educational goals and objectives as part of the evaluation process. The lack of discussion and theoretical analysis of the purposes of evaluation is consistent with the virtual absence of data on what educators at all levels believe the purposes of evaluation to be or of how they might prioritize multiple purposes.

The literature on the purposes of evaluation in schools and classrooms is a literature of lists and incidental notes produced by various commentators and researchers, some of whom have noted a purpose or two of evaluation by way of introduction to other issues. A synthesis of these lists and items produces the master list presented in Table 1.

Insert Table 1 About Here

Four generic functions appear in these statements of the purposes of evaluations and permit a reasonably parsimonius classification. These functions are certification, selection, direction, and motivation. Each represents a distinct purpose of the evaluation processes that occur in schools and classrooms.

Certification refers to the assurance that a student or program has attained a certain level of accomplishment or mastery. At the program level, certification typically involves some type of accreditation. At the individual level, certification might involve the

issuance of some sign of assurance such as a diploma or a recommendation for promotion.

Selection is the identification of suitable individuals, subgroups, and groups of individuals to be recommended for or permitted to enter or continue along certain educational and occupational paths. Evaluations are used to identify students for courses of study, programs, higher educational opportunities, and various levels of employment. At the program level, selection involves choices among competing programs for continuing public support. The expected outcome is the improvement of individual and program performance.

Direction refers to the use of evaluation processes to communicate to those being evaluated the specific desires of the evaluators. Evaluations provide key information to focus the attention of those being evaluated, whether they be the students in a classroom, or the teachers and administrators implementing an educational program. Such information may be criteria that communicate the appropriate emphases on tasks, or standards that communicate the desired level of performance.

Motivation entails involving those being evaluated in the tasks upon which the evaluation will be based. If the directing function or purpose of evaluations assures that individuals are aware of how they are expected to perform, the motivation function or purpose assures that individuals will be willing to commit the effort necessary to perform the task.

These four purposes of evaluation -- certification, selection, direction, and motivation -- have important effects on the other elements of evaluation processes. Although no data exists on the relative role of these four purposes in evaluation processes in schools and classrooms, most evaluation systems reflect at least some interest in each.

Student Tasks in Schools and Classrooms

The assignment of tasks to students is the beginning of the evaluation process in classrooms -- the student must first be given the responsibility for performing the task. While students generally work with a relatively stable set of tasks, the specific student tasks will constantly change if students are making the expected progress in their development.

The task assignment process consists of several distinctly different facets. Hackman (1969) defines a task as consisting of "...a stimulus complex and a set of instructions which specify what is to be done vis a vis the stimuli. The instructions indicate what operations are to be performed by the subject(s) with respect to the stimuli and/or what goal is to be achieved." Thus Hackman sees the task as consisting of stimulus materials, instructions about operations, and instructions about goals.

A similar approach is adopted by Dornbusch and Scott (1975), who distinguish between tasks assigned by delegations and those assigned by directives. The former involves specifying a goal and permitting

the performer to make at least some non-trivial decisions about how to attain that goal. The latter involves the selection of a path or set of activities which are then communicated with the expectation that the performer will carry out the prescribed course of action.

Doyle (1983:161) adds a third element to considerations of tasks in classrooms:

> The term "task" focuses attention on three aspects of students' work: (a) the products students are to formulate, such as an original essay or answers to a set of test questions; (b) the operations that are to be used to generate the product, such as memorizing a list of words or classifying examples of a concept; and (c) the "givens" or resources available to students while they are generating a product, such as a model of a finished essay supplied by the teacher or a fellow student. Academic tasks, in other words, are defined by the answers students are these answers.

Classrooms and schools are dominated by tasks. Doyle (1983:162) argues that tasks are crucial features of schools and classrooms, "that tasks form the basic treatment unit in classrooms" and that:

1. Students' academic work in school is defined by the academic tasks that are embedded in the content they encounter on a daily basis. Tasks regulate the selection of information and the choice of strategies for processing that information. Thus, "changing a subject's task changes the kind of event the subject experiences" (Jenkins, 1977:425).

2. Students will learn what a task leads them to do, that is, they will acquire information and operations that are necessary to accomplish the tasks they encounter (see Frase, 1972, 1975). In other words, accomplishing a task has two consequences. First, a person will acquire information--facts, concepts, principles, solutions--involved in the particular task that is accomplished. Second, a person will practice operations--memorizing, classifying, inferring, analyzing--used to obtain or produce information demanded by the task.

Because the nature of student tasks has such a pervasive influence on the classroom, it is important to understand the tasks which dominate the academic work of students. Doyle (1983:162-163) identifies four general types of academic tasks in classrooms:

1) memory tasks in which students are expected to recognize or reproduce information previously encountered (e.g., memorize a list of spelling words or lines from a poem);

2) procedural or routine tasks in which students are expected to apply a standardized and predictable formula or algorithm to generate answers (e.g., solve a set of subtraction problems);

3) comprehension or understanding tasks in which students are expected to (a) recognize transformed or paraphrased versions of information previously encountered, (b) apply procedures to new problems or decide from among several procedures those which are applicable to a particular problem (e.g., solve "word problems" in mathematics), or (c) draw inferences from previously encountered information or procedures (e.g., make predictions about a chemical reaction or devise an alternative formula for squaring a number);

4) opinion tasks in which students are expected to state a preference for something (e.g., select a favorite short story).

The academic tasks that dominate schools and classrooms have important implications for evaluation and control processes. Certain characteristics of academic tasks are particularly likely to affect the operation of the evaluation system in a classroom. For example, Dornbusch and Scott (1975:80) have suggested that tasks differ in predictability, that is, "the extent to which the performer has knowledge of which path is most likely to lead to success."<1> They argue that the greater the predictability of a task, the more likely that it will be assigned by a directive which specifies the path or procedures to be followed in executing the task. On the other hand, tasks that are low in predictiability will be

more likely to be assigned by a delegation which specifies a desired end state or goal and gives the performer the autonomy to make non-trivial decisions about how to attain that end state. Thus the task assignment that starts the evaluation process in motion would be likely to differ depending upon the nature of the task. Moreover, Dornbusch and Scott (1975) demonstrate that when tasks are predictable, performers prefer directives; when tasks are unpredictable, performers prefer delegations.

Doyle (1979) linked the nature of tasks to the evaluation process in terms of the ambiguity and risk associated with academic work in classrooms. He argued that because academic tasks in classrooms were performed in the context of an evaluation system, they were performed under conditions of varying ambiguity and risk. "Ambiguity refers to the extent to which a precise answer can be defined in advance or a precise formula for generating an answer is available...Risk refers to the stringency of the evaluative criteria a teacher uses and the likelihood that these criteria can be met on a given occasion" (Doyle, 1983:183). He classified understanding and opinion tasks as high in ambiguity and memory and routine tasks as low in ambiguity. Opinion tasks and certain memory tasks (i.e., those involving the reproduction of small amounts of material) and certain routine tasks (i.e., those requiring relatively simple algorithms) were classified as low in risk. Understanding tasks and other memory tasks (i.e., those involving the reproduction of large amounts of material) and other routine tasks (i.e., those involving complicated procedures) were classified as high in risk.

Academic tasks that are less predictable or that carry greater ambiguity and risk place greater demands on evaluation processes. Dornbusch and Scott (1975) and Thompson (1967) observe that when organizational goals are ambiguous as opposed to crystalized, performers may receive vague quality criteria. Such criteria often result in an evaluation process that is "...arbitrary and post hoc at every step, with the result that performers are unable to relate the performances to the evaluations received" (Dornbusch and Scott, 1975:258). When the relationship between procedures or operations and results or products is predictable, student performance can be evaluated by collecting information on the results or products. Indeed, such products can be designed for the convenience of the teacher as an evaluator. This is even more convenient when the tasks are low in ambiguity and thus have a clearly defined and precise product. Such tasks could be evaluated relatively easily no matter what the purpose of the evaluation -- certification, selection, direction, or motivation.

However, when tasks are low in predictability, the evaluation cannot rely solely upon inspection of a product or result if the purpose of the evaluation is to provide direction or enhance motivation. Inspection of results for an unpredictable task or a task high in ambiguity does not provide sufficient information for an evaluator to use to help students improve their performance. Moreover, evaluations of unpredictable tasks based on results or products provide no information for the evaluator to determine the effort and performance that led to the product. As a result, the

evaluator will not be able to structure the evaluation to maintain or enhance the motivation of the student. This problem is compounded by the fact that academic tasks generally involve mental processes that are not readily visible to teachers in classrooms (Natriello and Dornbusch, 1984).

Two processes appear to be set in motion by the strain that unpredictable and ambiguous tasks place on evaluation. First, there is a tendency to avoid unpredictable or ambiguous tasks in schools and classrooms. Doyle (1983) reviews studies by Davis and McKnight (1976) and Wilson (1976) which suggest that students resist the shift from routine or procedural tasks to understanding tasks in classrooms. After trying to make such a shift in a mathematics class, Davis and McKnight (1976:282) commented that "it is no longer a mystery why so many teachers and so many textbooks present ninth grade algebra as a rote algorithymic subject. The pressure on you to do exactly that is formidable." Besides resisting the introduction of less predictable tasks, students also attempt to renegotiate assigned tasks so that they are more predictable by soliciting more information from the teacher on the specifics of the performance and results desired.

Teachers may also devote less attention to less predictable tasks when evaluating students. Natriello and Dornbusch (1984) demonstrate that teachers and administrators present students with more frequent and more challenging evaluations for behavior tasks which are conceived of as more predictable than for academic tasks which

are conceived of as less predictable. This same reasoning is used by Holmes (1978) to explain why schools and teachers are less likely to take seriously the evaluation of students in the affective domain, where tasks are conceived of as even less predictable than in the academic domain.

A second process may be set in motion by the strain that unpredictable tasks place on evaluation systems -- the tendency to structure evaluation activities as if the tasks being evaluated are predictable and unambiguous. In a study of three reading curricula, Armbruster, Stevens, and Rosenshine (1977) found that although the texts emphasized comprehension and interpretation skills, the tests solicited factual information from students based on the ability to locate information in the text. Treating tasks as if they are predictable simplifies the evaluation process.

Although types of tasks differ in ways that affect evaluation processes in classrooms and schools, all academic tasks are complex. Reviewing recent research on tasks and cognitive development, Doyle (1983:173) points out that "In sum, school tasks, even at the level of basic skills, are inherently complex for all students. This complexity is much more severe, however, for young students and those who lack either the information or the skills required to understand tasks, process information in specific ways, or decide when to use the strategies they possess." Such complexity carries important implications not only for evaluation and control systems (Dornbusch and Scott, 1975), but also for the structure of work groups and

organizations (Scott, 1981). More complex tasks require more sophisticated evaluation processes to assess student performances accurately.

Setting Criteria

The assignment of a task communicates to a student that he or she is responsible for performing that task. However, evaluators are generally interested in more than the performance and completion of tasks (Dornbusch and Scott, 1975); they are also concerned with certain properties and levels of the performance and of the final product. Thus, in addition to the task assignment, the evaluation system requires the setting of criteria and standards for a task.

Considerable confusion surrounds these issues in the evaluation of student performance in schools and classrooms. One type of confusion -- the failure to clearly distinguish between criteria and standards -- was introduced in the literature on criterion-referenced testing and has now become pervasive among researchers and practitioners alike, no doubt testimony to the effectiveness of courses in tests and measurement.

Glass (1978) traces the use of the term "standard" in the work of Mager (1962) and Popham (1973) on instructional objectives, Bloom (1968) on mastery learning, and Tyler (1973) on the role of testing in assessment programs. In each case "standard" is used to refer to a level of acceptable performance in behavioral terms. He next considers the work of Glaser (1963) on criterion-referenced tests, work

in which Glaser assumed that there were continua of attainment levels along which student performance could be described. Finally, he argues that

> Glaser's use of the word "criterion" with its colloquial meaning of "standard," the simultaneous publication of Mager's rather simple notions of performance standards, and Popham's mixing of Glaser and Mager in the same pot combined to create the impression that the "criterion" in criterion-referenced testing was not the behavioral scale articulated to a test and elaborating the meaning of the scores, but rather that the "criterion" was the cut-off score, the division between pass and fail, or competence and incompetence. This interpretation of the word "criterion" is evident in the informal conversation of both educators and measurement specialists. This meaning is intended when people speak, as they do now habitually, of "setting the criterion on a criterion-referenced test or test item." (Glass, 1978:241)

The continuing confusion of the terms "criteria" and "standards" makes it particularly important to distinguish them in considering their role in evaluation systems. Criteria refer to the properties of the task that should be taken into account in making the evaluation (Dornbusch and Scott, 1975:138). A standard, on the other hand, refers to the evaluative scale whose ...

> intervals constitute degrees of acceptability or preference, the scale typically ranging from low scores indicating 'totally unacceptable' values at one end of the continuum to high scores indicating 'highly acceptable' or perhaps 'exceptional' values at the other end. A standard may consist of a single point on the evaluative scale separating acceptable from unacceptable values. More typically, however, a standard consists of a set of points distinguishing various levels of acceptability or non-acceptability. In addition to the scale itself, the standard also includes a set of rules to transform values on the performance dimension into scores on the evaluative scale (Dornbusch and Scott, 1975:140).

A second type of confusion, also rooted in the traditional treatment of behavioral objectives, involves the failure to distinguish

the dual aspects of a standard -- the component related to the levels of the important properties of the assigned task (i.e., the criterion levels), and the component related to the collection of information on the performance dimensions (i.e., the sampling process). Discussions of behavioral objectives (e.g., Gronlund, 1971; Krathwohl and Payne, 1971; Brown, 1970; Lindvall, 1961; Remmers, Gage, and Rummel, 1960; Lien, 1967; Ahmann and Glock, 1967) typically present levels of objectives that range from general objectives (criteria) to specific objectives which have the desired student behaviors clearly identified (indicators). Such presentations unintentionally confuse the properties that are of interest to the teacher or evaluator with the evidence of student performance in terms of those properties and the standards for performance.

Such melding of criteria, standards, and indicators may make the development of learning objectives more concretely understandable to teachers, but it also locks them into an overly empiricist conception of these relationships. The specific objectives (indicators) are interpreted as being in one-to-one correspondence with the general objectives (criteria).<2> Under such circumstances, the indicators of student performance can take on the role and importance of the criteria themselves. Levine (1976) observes this process occurring in achievement testing in schools, both in the writing of experts who argue that achievement tests are absolute criteria in themselves (e.g., Lindquist, 1969) and in cases where the testing program has dictated school policy (Levine and Levine, 1970). To avoid such an empiricist trap, the present conceptual scheme distin-

guishes among criteria, standards, and the indicators of performance.

Serious discussions of the criteria properly associated with student academic tasks tend to be specific to various content or curricular areas. For example, Doyle (1983) reviews Culler's (1980) analysis of the criteria involved in competence in literature and Fredriksen and Dominic's (1981) analysis of the criteria involved in the composing process. Because criteria for the evaluation of student performance must identify the important properties of various student tasks, the criteria themselves must be specific to the tasks if they are to have meaning in the context of the work of students. Yet discussions of the evaluation of student work typically pay little attention to the specific tasks being evaluated. Rather, an evaluative technique is applied which may or may not be appropriate for the tasks in question. As noted earlier, the application of such techniques may then transform the nature of the tasks to conform to the evaluation process.

While there is little discussion of task-specific criteria for evaluation in the evaluation literature, attention has been devoted to the types of criteria employed in the evaluation process. The achievement of students in a subject is generally accepted as the one criterion common to all evaluation systems in schools and classrooms (Brown, 1970). The appropriateness of using achievement criteria is seldom discussed, although increased attention is being paid to determining whether the evaluation process is linked to the

instructional process (Linn, 1983; Rudman, Kelly, Wanous, Mehrens, Clark and Porter, 1980). The latter is generally accomplished by matching the testing procedures to the goals and objectives of instruction,<3> so that students are not subjected to an evaluation process that involves things not covered in the instructional program -- a problem for both groups of students and for individual students (Natriello, 1982).

Types of criteria other than achievement criteria enter into evaluation processes in schools and classrooms, but there is little agreement as to which of these are appropriate. For example, Thorndike (1967:762) notes that:

> In practice, certainly many other considerations than that of pure competence do enter into marks. Such factors enter in as (1) industry and effort--i.e., completing all assigned work and even doing optional work for "extra credit" (a kind of educational bribe); (2) frequent and active participation in class discussion; (3) neatness in written work and mechanical correctness in such areas as spelling and grammar; and (4) personal agreeableness, attractiveness, cleanliness, and docility. To some extent and by some instructors, certain of these features would be endorsed as legitimate influences on a mark. Others would more uniformly be accepted as extraneous influences, to be minimized as far as possible.

Holmes (1978), observes that criteria related to behavior and effort, and particularly criteria such as politeness, conformity, and perserverance -- those things which make it possible for the organization to operate -- often covertly enter into the evaluation process. He argues that more formal attempts should be made to include criteria from the affective domain (such as attitudes, values, and moral reasoning) in student evaluations. Brown and Craig (1977) value these other criteria, but reject the notion that

they can be incorporated into systems for the evaluation of students in schools and classrooms.

Several studies have examined the use of criteria other than achievement in evaluation systems for students. Schunk (1983) reports on an experiment in which students were subjected to three types of evaluation systems -- one in which they received rewards for their actual performance, one in which they received rewards for simply participating (i.e., honoring the task assignment), and one in which they received no rewards. While the performance contingent reward system led to the highest levels of achievement, the system of rewards for participation showed no benefit over the no-rewards system. Salganik (1982) reports on a system in which students were evaluated on three criteria -- achievement, effort, and conduct. Although the correlations among these three types of criteria were high, in those cases in which there were discrepant evaluations among the three criteria, the evaluations based on student effort seemed to have some positive motivational effects on low-achieving students. Weiner (1979), reviewing research on attributions and motivation, found that evaluators placed greater importance on effort than on ability in determining reward and punishment under conditions in which performance was held constant. Natriello and McPartland (1987), in a national survey of secondary school teachers, found that student effort was "very important" or "extremely important" in the evaluation process of over 70% of the teachers.

Types of criteria other than those related to achievement clearly play a part in the evaluation of students in schools and classrooms. Additional research should provide descriptions of these non-achievement criteria and the ways in which they are used by teachers. Because teachers report the use of multiple criteria (Natriello and McPartland, 1987) and because students are assigned multiple tasks in classrooms, some attention should be paid to the relative weight assigned to various tasks and criteria in arriving at final evaluations (Brown, 1970; Dornbusch and Scott, 1975).

Standards

The standards used in the evaluation of students have received considerable attention from both the public at large and educational researchers. Recently, there have been renewed calls for higher standards in U.S. schools (National Commission on Excellence in Education, 1983). Calling for "higher" standards requires only the assumptions that current standards are too low, and that "higher" standards will somehow lead to better educational outcomes -- assumptions for which there is reasonable evidence at least for some groups of students (Natriello and Dornbusch, 1984; McDill, Natriello, and Pallas, 1985). But calls for "higher" standards rely upon current standards as a point of reference and thereby avoid a key area of controversy in the evaluation of students.

Researchers and practitioners have produced a considerable body of work which examines the appropriate reference point for establishing standards for evaluating students. This work, which typi-

cally considers various systems that might be employed for setting standards, appears to grow out of a fundamental dilemma faced by schools in evaluating student performance. This dilemma is clearly described by Bidwell (1965:973) in discussing the school as a formal organization. First, he notes that in order to produce a uniform product, schools and teachers engage in the "universalistic and thus uniform assessment of student accomplishment." Teachers present organizational standards to students and evaluate their performance in terms of those standards. However, presenting standards to students and demanding their compliance may not be enough to promote learning, in view of the fact that students may not see schools as relevant to their immediate interests. Citing Waller (1932), Bidwell (1965:979) points out that "motivation to learn...is very largely a product of a close, warm relation between teacher and student," and suggests that the nature of school organizations requires teachers to exhibit instances of universalistic assessment, as well as instances of more diffuse responsiveness, in order to be effective with students. Similarly, Varenne and Kelly (1976) see the school as caught by the paradox characteristic of American culture which incorporates both the belief in the equal endowment of all with regard to certain inalienable rights and their unequal endowment with regard to individual capacities. This paradox requires the school to utilize universal criteria for evaluation and rewards that are tailored to individual performance.

Research and commentary on appropriate standards for the evaluation of student performance in schools and classrooms examine the

struggle to accomodate both universalism and individualism in a single system. Out of this struggle have emerged three types of standards: those set in reference to the criterion level of a group, those set in reference to some absolute criterion level, and those set in reference to the previous criterion level of an individual (Wise and Newman, 1975; Rheinberg, 1983; Thorndike, 1969). Discussions of standards for the evaluation of students revolve around the advantages and disadvantages of employing each of these three types.

Norm- or group-referenced standards have been criticized by educators and social scientists alike, perhaps because they have been in widespread use for such a long time. Terwilliger (1978) refers to the use of norm-referenced standards as "norm-referenced grading" and specifies four variations. The most commonly discussed approach is the use of a normal curve with a specific class, a practice that Terwilliger traces to the "scientific movement" in education in the 1930's, an observation borne out by Crooks' 1933 account of then current thinking on grading. In this method, teachers use the test scores of students in their class to create the normal or bell-shaped curve. This method is typically referred to as "grading on the curve" (Bresee, 1976). A second method relies on the same normal curve but includes the evaluative scores of a larger group of students beyond the immediate class, (e.g., all of the students receiving similar instruction currently or in the recent past). A third variation is restricted to the immediate class, but assigns grades based on a distribution other than the normal curve. A fourth variation assigns grades using a distribution other than the normal curve and uses a reference group beyond the immediate class.

Additional varieties of norm-referenced standards have been identified by Michaels (1977) and Slavin (1977) while discussing classroom reward structures. Michaels (1977) defines "individual competition" as a reward structure in which grades are assigned to students based on their performances relative to those of classmates and "group competition" as a reward structure in which grades are differentially allocated to groups according to their relative performance. Slavin (1977) designates similar reward structures as "competitive reward structures" and "group competition," respectively. Both analyses point out that norm-referenced standards can apply to different levels in a system -- individual students, groups of students, programs, schools, etc.

Terwilliger (1978) links norm-referenced standards to what he terms the pragmatic philosophy, a viewpoint primarily concerned with practical choices and the consequences of such choices. An evaluation system which differentiates among individual students is optimal for identifying the available choices and their consequences. Thus, norm-referenced standards would appear to serve the purpose of selection identified earlier. Rheinberg (1983) links norm-referenced standards to the rationale of psychological testing and the associated concerns for objectivity, reliability, and validity of a teacher's grading process. He notes that "Perhaps because of this orientation towards psychological testing theory, an implicit assumption was perpetuated: 'Correct' evaluation of academic achievement has to be based on social comparison between students, leading to a normal distribution of grades." (Rheinberg, 1983:185).

Levine (1976:233-234) explains how the interests of educational psychologists in producing distributions of scores amenable to statistical analysis overrode the interests of teachers who would have preferred scales which enabled them to see where students were and where they had to go, and who would have preferred not to harm students by using national norm-referenced standards which placed half of them below the national standards.

The extent to which teachers actually use norm-referenced standards has received too little attention. Rudman, Kelly, Wanous, Mehrens, Clark, and Porter (1980:32), after reviewing studies which described teacher testing preferences (i.e., Yeh, 1978; Olejnik, 1979; O'Regan, et al., 1979; Nearine, 1970), conclude that "Those studies that were descriptive tended to show a preference for norm-referenced tests and the standard scores in which the results are couched." On the other hand, in the national survey reported on by Natriello and McPartland (1987), secondary school teachers rated norm-referenced standards as less important than either criterion-referenced standards or individually-referenced standards in determining student grades. In addition, Gullickson's (1982) survey of South Dakota teachers revealed that only 10% of the respondents reported grading on a curve. Rheinberg (1983) notes that teachers who preferred norm-referenced standards tended to organize classroom tasks so that all students engaged in uniform tasks to facilitate comparisons and to view student achievement differences as very clear and stable properties of students. Although further evidence is necessary before reaching conclusions, it may be that teachers

seek to use formal achievement tests with norm-referenced standards to balance their own criterion-referenced and individually-referenced standards in the classroom.

Although little evidence exists on the extent to which norm referenced standards are actually used in schools and classrooms, critiques of the practice abound. General critiques are provided by Bresee (1976) and Deutsch (1979). Bresee (1976) lists a series of problems with such standards: (1) the necessity of producing a normal distribution of grades conflicts with the goal of having teachers produce improvement in all students in a class; (2) the distortion of the curriculum as teachers seek to diversify instructional objectives to produce a range of achievement in a class; (3) the diversion of student attention from the task at hand to the performance of other students; and (4) the introduction of false competition because achievement is not really in limited supply. To these Deutsch (1979) adds: (1) the distortion of the testing process so that tests take the form of contests in which all performers participate under uniform conditions; (2) the lack of rewards created by the artificial scarcity of good grades that is likely to impede the development of students' sense of their own value; and (3) the encouragement of competition which may be counterproductive for tasks requiring cooperation and communication.

Implicit criticisms of norm-referenced standards have come from advocates of criterion-referenced testing<4> (Glaser, 1963), who tend to point out how unsuited such standards are for providing

insight into the effectiveness of educational treatments or programs; and from advocates of individually-referenced standards (Beady and Slavin, 1981), who decry the deficiencies of relative standards for providing direction and motivation for certain students. Thus, the critiques of norm-referenced or relative standards center around the application of those standards to purposes such as accreditation, direction, and motivation, for which they are ill-suited.

Criterion referenced or absolute standards have enjoyed a great deal of attention due to the criterion-referenced testing movement. Glass (1978) observes that contemporary educational movements for accountability, mastery learning, assessment, competency-based education, and minimal competence graduation requirements have received increasing attention. Each of these approaches relies on some absolute set of standards.

Terwilliger (1978) notes the forms that the use of absolute standards in the classroom can take. He identifies the "percent-correct system" as an approach "in which 100 represents a perfect performance and some arbitrarily designated value (e.g., 70) represents the minimal 'passing' score. If letter grades are employed, grades are defined in terms of specified ranges on the percent-correct scale, e.g., A=95-100, B=87-94, C=78-86, D=70-79, F=69 or below." (Terwilliger, 1977:31). A second approach attempts to build specific meaning into criterion-referenced systems by specifying the minimal level of performance that is acceptable. A third, more lim-

ited, approach to absolute standards focuses on the quantity of a certain task that is completed by a student.

Discussions of reward structures by Michaels (1977) and Slavin (1977) suggest additional approaches to absolute standards. Michaels (1977) uses the term "individual reward contingencies" to describe a reward structure in which the performance of individual students is compared to a previously established standard. He uses the term "group reward contingencies" to describe a reward structure in which the performance of each group is independently compared to a previously established standard. Slavin (1977) uses the term "independent reward structure" to describe a reward structure in which the probability of a student's receiving a reward is unrelated to the probability of any other student receiving a reward (as when the performance of individual students is compared with a fixed standard). He uses the term "group contingencies" to describe a situation in which the group is evaluated against a fixed standard.

Terwilliger (1978:23) associates absolute standards with the behaviorist perspective on education which argues that "the optimal conditions for learning require a highly structured individualized approach in which materials are presented in relatively discrete units." and "...stresses the need for identifying in advance: 1) the precise objectives of instruction, 2) the exact instructional objectives to be employed, and 3) the specification of the criteria used for judging whether the objectives have been attained."

Few studies have examined the extent to which teachers actually use absolute criteria in evaluating student performance. Approximately three-fourths of the teachers in the national sample examined by Natriello and McPartland (1987) reported that absolute standards for achievement were "very important" or "extremely important" in arriving at a final grade for students in their classes. This is consistent with Gullickson's (1982) finding that 78% of teachers in his South Dakota sample reported using some kind of criterion-referenced grading scheme. However, Rudman, et al. (1980) point out that although commentary on test use suggests that teachers prefer criterion-referenced tests over norm-referenced tests, descriptive studies show the opposite and only 35% of the teachers surveyed by Beck and Stetz (1979) favored increased use of criterion-referenced tests.

Discussions of absolute standards have paid considerable attention to various methods for arriving at mastery, competency levels, or cutting scores (Berk, 1976; Hambleton, Swaminathan, Algina, and Coulson, 1978; Meskauskas, 1976; Nedelsky, 1954). Glass (1978) and Burton (1978) review various methods for determining where to set a mastery level on a continuum and conclude that standards must be set arbitrarily. Shephard (1976) concludes that current methods of setting absolute standards all reduce to a form of norm-referenced standards. The inability to set standards by other than arbitrary means causes Glass to reject absolute standards on standardized tests and argue for the use of improvement as a basis for evaluation. Scriven (1978) provides a counter perspective which argues

that absolute standards are not totally arbitrary and may still be employed in minimal competency testing. However, these arguments serve to underscore the fact that absolute standards are problematic, particularly in cases such as statewide testing programs in which decisions about standards are removed from the informed professional opinion of the teacher (Burton, 1978).

Individually-referenced or self-referenced standards are based on comparing a student's current performance with some other feature of the student. Terwilliger (1978) distinguishes two forms of self-evaluation, comparing current performance with earlier performance and assessing growth, and comparing current performance to a student's ability. Terwilliger views the use of self-referenced standards as an attempt to "recognize individual differences, reward effort and generally provide an environment which fosters interest and motivation" (Terwilliger, 1978:32). He associates it with the humanist view of education which is concerned with "the values, interests, and dignity of each individual student as a human being" (Terwilliger, 1978:24).

Rheinberg (1983) links self-referenced or individually-referenced standards to the work of European educational theorists such as Herbart and Pestalozzi. He provides quotes from each -- "The teachers does not compare his student with others but with the student himself" (Herbart, 1831:10) and "I was patient with the slowest learner; but if one of the students did something worse than before I was harsh" (Pestalozzi, 1807:426) -- to illustrate their longitudinal perspective on individual standards for evaluation.

It is unclear to what extent teachers employ individually-referenced standards in evaluating student performance, though such standards do play a role in evaluation processes in classrooms. Rudman et al. (1980) found that 77% of the teachers in the study by Beck and Stetz (1979) favored using standardized test data to measure student growth. Natriello and McPartland (1987) report that about three-fourths of the teachers in their national sample rated self-referenced standards as "very important" or "extremely important" in arriving at final grades. Rheinberg (1983) found that teachers who did report a preference for self-referenced standards tended to individualize classroom tasks and to view student achievement as flexible and present-oriented. Finally, a number of investigators have developed programs to establish individually-referenced evaluation processes in classrooms (Hansen, 1977; Ready, Slavin, and Fennessey, 1981). The preponderance of work on individually-referenced standards suggests that they are particularly appropriate to the purposes of motivation and direction noted earlier.

Collecting Information on Student Performance

In a rationally ordered evaluation process, once decisions have been made about the purposes of evaluation, the tasks, the criteria, and standards, an evaluator would be in a good position to consider the appropriate strategy for collected information on student performance. The collection of such information requires a sampling process because it would be impractical if not impossible to collect total information on student performance. Most of the important

decisions about the collection of performance information thus involve sampling decisions to insure that the information collected provides a valid and reliable estimate of performance appropriate to the purposes, tasks, criteria, and standards that have been already determined. Of course, in many instances the evaluation process is not rationally ordered and decisions on the collection of information on student performance seem poorly articulated with the purposes, tasks, criteria. and standards.

The dominant technique for collecting information on student performance is some form of testing. This is true at the federal and state levels, where formal assessment programs have proliferated in recent years; at the district level, where school administrators and local boards of education have become increasingly concerned with the performance of the system; and only slightly less true at the classroom level, where teachers rely on their own tests for a number of reasons (Herman and Dorr-Bremme, 1984; Rudman, et al., 1980).

A number of analysts have contributed important observations about the relationship between testing practices and the purposes, tasks, criteria, and standards for the evaluation of students. Deutsch (1979) argues that the structure of most testing situations is dictated by the prevailing purpose of evaluation (selection) and the types of standards utilized (norm-referenced). He notes that:

> The social context of most educational measurement is that of a contest in which students are measured primarily in comparison with one another rather than in terms of objective accomplishment. If educational measurement is not mainly in the form of a contest, why are students often asked to reveal their knowledge and skills in carefully regulated test situations designed to be as uniform as possible in time, atmosphere, and

> conditions for all students. Individuals vary enormously in terms of the amount of time they need and the kind of atmosphere and circumstances that facilitate or hinder the expression of their knowledge and skills; it is only the comparison of students with one another that requires measures of educational achievement to take the form of contests (Deutsch, 1979:394).

Deutsch goes on to describe the damaging effects of norm-referenced standards for individual students and advocates an evaluation system that would provide individualized, particularistic feedback to students to foster their development. Thus, his objection to the typical testing situation is rooted in a rejection of the selective purpose and the norm-referenced standards that characterize much evaluation in schools and classrooms in favor of individually-referenced standards that might contribute to student motivation.

Others have also rejected testing strategies rooted in norm-referenced standards while advocating a criterion-referenced approach. These discussions typically object to the selection of items for standardized tests, which is a sampling strategy in itself. Popham and Husek (1969) point out that the appropriate strategy for sampling items for tests when the standards are norm-referenced is to select items which maximize the variability of performance among the individuals taking the test. Hambleton, et al. (1978) observe that criterion-referenced tests are not constructed to maximize the variability of test scores, so the resulting distributions will tend to be homogeneous. They go on to note that norm-referenced tests are sometimes used to make criterion-referenced measurements and criterion-referenced tests are sometimes used to make norm-referenced mea-

surements, but that neither strategy is particularly satisfactory. Both these authors and others (Glaser, 1963) base their arguements for criterion-referenced tests on the inappropriateness of using norm-referenced tests for purposes of certification.

The purposes for which tests are used in schools and classrooms have recently been examined by Herman and Dorr-Bremme (1984) in their national survey of administrators and teachers. Table 2, adapted from their technical report (Herman and Dorr-Bremme, 1984:43), presents the percentages of principals reporting that test results and other kinds of information are crucial or important for particular purposes in the school.

Insert Table 2 About Here

We can compare the ratings for the types of formal tests listed in the first three columns of the table in terms of two purposes -- selection and certification -- which are included in our four category system developed earlier. For assigning students to classes, an example of selection, norm-referenced tests are rated as important more often than either minimum competency tests or district objectives-based tests. This is true for both elementary and secondary principals and is consistent with what is generally understood to be the best use for norm-referenced tests. For student promotion decisions, an example of certification, minimum competency tests are more often rated as important for this purpose at the secondary

level (as might be expected), but norm-referenced tests are seen as important more often at the elementary level (though only slightly more so than district objectives-based tests). But a second trend overshadows these patterns of responses regarding formal tests. The results of teachers' classroom testing are rated as important more often than the results of any of the three formal tests, and teachers' opinions, judgments, and recommendations carry more influence than any of the test results. Thus, the source of the information (i.e., its generation within the school) appears to be more important than the type of information for influencing decisions. These patterns of results are confirmed in teacher responses to questions regarding the use of various sources of information for making classroom decisions (Herman and Dorr-Bremme, 1984:48-55).

The relationship between academic tasks, the criteria for defining and judging them, and testing have also been the subject of considerable discussion and inquiry, typically under the rubric of the relationship between teaching and testing or integrating instruction and assessment. Improving the relationship between what is tested and what is taught is a major issue in the improvement of testing in U.S. schools (National Institute of Education, 1979). The poor fit of tests to the academic tasks assigned to students has concerned educators in particular subject areas (such as social studies) which are often outside the basic areas where most test development activity is concentrated (Rimmington, 1977), as well as researchers, who worry that differences in the degree to which tests correspond to academic tasks will produce biased evaluations of educational programs (Leinhardt and Seewald, 1981).

Rudman, et al. (1980) review a wide range of information on the integration of assessment with instruction and find few careful analyses of the relationship between the nature of academic tasks in classrooms and the content of tests. Leinhardt and Seewald (1981) note that analyses of the relationship between teaching and testing are expensive and time consuming. They review a number of techniques for analyzing the correspondence or overlap between teaching and testing and conclude that, although all such analyses are complex, those involving elementary education in the basic skills are somewhat easier to do. This suggests that basic skills testing will be the area in which most care will be taken to match testing strategies to the nature of academic tasks.

The practicality of the relatively less complex basic skills tests also appears to affect the nature of tasks in schools. In the national sample of administrators and teachers in the Herman and Dorr-Bremme (1984) study, respondents in both groups reported that increased testing has resulted in more instruction in the basic skills. Nearly three-quarters of the principals report that as a result of testing programs, more instructional time is being devoted to the basic skill subjects of reading/English and mathematics. Among teachers, 88% of the elementary teachers, 84% of high school English teachers, and 74% of high school math teachers reported that instruction in the basic skills was consuming a substantially greater portion of the school's educational resources. Moreover, the impact of testing programs in promoting greater attention to the basic skills appears to be greater among schools serving students of lower socioeconomic status.

Several recent studies provide some basic descriptive information on the use of tests in schools and classrooms. Gullickson (1982) surveyed teachers in South Dakota about their testing practices. Responses revealed that 89% of elementary teachers and 99% of secondary teachers relied on some kind of testing, and most tested at least weekly (95%) or biweekly (98%). Although teachers reported using a variety of testing techniques, "...only teacher-made objective tests played a major evaluative role across all grade levels and curricular areas" (Gullickson, 1982:3). Further, "...teachers reported teacher-made objective tests as having the greatest role, essay tests as having the second largest role, followed by standardized objective tests and oral quizzes. Of the four, objective tests received much higher ratings than did all of the other three. Essay tests received high ratings at the secondary level but very low ratings at the elementary level" (Gullickson, 1982:4). Despite the predominance of objective tests, teachers reported believing that essay tests provide a better measure of learning, particularly for higher cognitive levels (Gullickson, 1984). Finally, teachers agreed that tests should not be the only basis for grading students, but about half of the respondents reported that tests do provide the primary basis for arriving at grades (Gullickson, 1984).

The conditions of testing reported by teachers in Gullickson's study confirm Deutsch's (1979) observation about uniformity to facilitate comparisons among students. Gullickson (1982:8) reported that:

> Testing appears to be a formal, constrained situation in which students expect to be graded. Virtually all teachers (99%) do

not allow student interaction during the testing process. A substantial percentage do not even allow students to ask questions of the teacher. In addition students are constrained in their use of support material. Seventy-nine percent of the teachers do not allow students to use their textbook, notes, etc., in completing a test.

Despite the controlled conditions under which teachers administer tests, Gullickson's (1982) analysis raises a number of troubling questions regarding teachers' testing practices:

> First, in the preparation of tests, short answer and matching items are the most popular items of choice. Both types tend to be limited to lower cognitive level, i.e., knowledge level, assessment (Hopkins and Stanley, 1981). Thus tests probably assess only lower cognitive level understandings. Second, while the large majority of teachers reuse items, few teachers take the time or make the effort to systematically improve their items. This is suggested by the minimal amount of time given to test analysis (barely enough to score and grade tests) and by the minimal use of test statistics. As a direct result, test item improvement must be done on a very ad hoc and subjective basis. Third, teachers appear to misuse criterion-referenced tests. On the surface teachers' advocacy of criterion-referenced testing would indicate evidence of a firm criterion-referenced testing foundation. However, even if teachers clearly define their test domain -- a topic not addressed in this survey -- they clearly do not address quality of items in a manner which would insure their items function as desired. Most reuse their items but without careful item analysis. Thus, criteria established by teachers are both artificial and subjective. For without knowing how items function, it is not possible to accurately set criterion levels for student performance (Gullickson, 1982:13-14).

Herman and Door-Bremme's (1984) national survey of administrators and teachers also provides insight into the basic test use patterns of teachers. Survey responses indicated that elementary students spend about four percent of the average instructional time devoted to reading and about seven percent of the average instructional time devoted to mathematics taking tests. These elementary students take a reading test and a math test about once every eight days. About

half of this time is spent on tests mandated by the district or the state. Secondary school students appear to spend more time taking tests. A typical tenth grade student spends about 13% of the average instructional time in English completing tests and about 12% of the average instructional time in mathematics completing tests. These high school students take an English test and a math test every three-to four days. About one-fourth of this time is devoted to tests mandated by the district or the state.

As noted earlier, both teachers and adminstrators see teacher-made tests as more important sources of information than district and state-mandated tests for making a variety of decisions in schools and classrooms. In view of the importance accorded teacher-made tests, Herman and Door-Bremme (1984) review some of the same concerns about the quality of teacher-made tests raised by Gullickson (1982). They write that:

> Recent research also indicates that teachers remain poorly prepared in assessment (Rudman, et al., 1980; Woellner, 1979; Yeh, et al., 1981). And as CSE's survey indicates, in-service training does little to fill the gap. Only about one-fifth of the teachers responding received staff development related to selection and construction of good tests or in the use of test results to improve instruction...In a recent review of teacher-made tests, Fleming and Chambers (1983) found that teachers write more questions of the short answer kind than of any other type; they rarely devise essay examinations. For the most part, too, the tests reviewed required students to recall facts and terms. Questions requiring learners to translate, apply, or otherwise use knowledge were rare. Furthermore, Fleming and Chambers discovered a "general tendency" to omit test directions, to use illegible test copies, and "to omit the point values to be assigned to test questions." Herman and Door-Bremme (1984:144).

These reservations about the quality of teacher made tests are consistent with the results of Natriello's (1982) interview study of teachers in four high schools. The interviews revealed that teachers varied greatly in their approaches to testing and evaluation, and many teachers lacked a well articulated approach to the evaluation of student performance in the classroom.

Although most research and commentary on the collection of information on student performance has centered on testing, alternative collection methods have been discussed and are used by teachers. Gaston (1976) observes that student behavior under testing conditions often fails to reflect tasks in the affective domain. He suggests alternatives to collect information about student attitudes and behavior, such as monitoring of students' unassigned reading in the library and listening to student conversations as students leave the classroom. Heller (1978) suggests alternatives to standardized reading tests such as the use of reading materials from popular magazines, fables, and poems. Solo (1977) explains how alternatives such as anecdotal records and collections of students' daily work may be used to provide insight into student performance. Herman and Door-Bremme (1984) note a variety of techniques used by teachers to collect information on student performance, including routine class and homework assignments, classroom interaction during question and answer sessions, recitations, discussions, oral reading, problem-solving at the chalkboard, special projects, presentations, and reports.

The national survey by Herman and Door-Bremme (1984) revealed that the teacher's own observations and classwork are more important than any type of testing for providing information for classroom decision making, and that teachers' opinions, judgments, and recommendations are more important than any type of testing in school decision making. Although such practices appear to broaden the base of information on student performance, there are serious questions about quality. Reviewing the literature on teachers' collection of information on student performance other than that supplied by tests, Rudman, et al. (1980:58) conclude that:

> Teachers' perceptions of students' behavior is stable and not much influenced by data when the new information seems to contradict what they have observed (Pedulla, Airasian, Madaus, and Kellaghan, 1977; Morine-Dershimer, 1979; Sorotzkin, Fleming, and Anttonen, 1974; Beggs, Mayer and Lewis, 1972)... In contrast to teachers' perceptions of their students' test scores there is some evidence that teachers' reporting of their students' classroom interpersonal behavior is neither stable nor accurate (Elmore and Beggs, 1972; Barnhard, Zimbardo, and Sarason, 1968; Openshaw, 1967; Feshbach, 1969; Tolor, Scarpetti, and Lane, 1967).
>
> Teachers seem not to be accurate observers of pupils' academic behavior. Several examples in the literature illustrate teachers' observations of oral reading by their pupils. Regardless of the amount of training or experience, teachers appear to make poor judges of the oral reading behaviors of students. (Ladd, 1961; Page and Carlson, 1973; Allington, 1978).

Thus, there is no shortage of serious questions about the use of tests and alternative methods for collecting information on the academic performance of students.

Appraising Student Performance

Appraising performance in a well developed evaluation system involves comparing the information collected on student performance

on assigned tasks with the criteria and standards previously established for those tasks. But even in a well articulated evaluation system, evaluators are expected to exercise judgment and discretion. As Dornbusch and Scott (1975) observe:

> The application of standards in specific situations is rarely a simple or straightforward procedure. It requires judgment with respect to the comparability of the performance situation and the situations for which the standards are considered applicable. Similar kinds of judgments are required in employing the specified property weights in combining scores to produce a performance evaluation. In short, appraisal is seldom a mechanical procedure. Moreover, task appraisal entails deciding how to interpret a low or high performance score. Accurately appraising a task performance requires knowledge of extenuating circumstances, whether it be the inexperience of the task performer, the lack of facilities, or assistance received from a more skilled co-worker. Such information is of critical importance in determining what, if any, message is to be communicated to the performer concerning the quality of his or her task performance (Dornbusch and Scott, 1975:143).

For some reason, the exercise of discretion that is expected of most evaluators is typically not expected of teachers by researchers who study the appraisal process. Indeed, the assumption has been that teacher appraisals which vary from the results of standardized tests of student performance are somehow flawed. Much of the literature on the appraisal process in the evaluation of student performance has focused on deviations of the appraisals from results of standardized tests. Such deviations are often characterized as teacher bias. The same perspective has been advanced by others to criticize standardized tests themselves, despite evidence that the major tests are not biased (Arnold, 1983).

Studies of teacher bias in appraising student performance have examined the effects of student characteristics on teacher apprais-

als of performance. After an extensive review of this literature, Natriello and Dornbusch (1984) concluded that four major problems with these studies limit the quality of the conclusions that might be drawn from them. First, although the literature suggests that certain groups of youngsters are more likely to be impeded academically by unsound teacher appraisals, the connection between teacher behaviors or attitudes and student achievement is assumed rather than documented. Second, these studies have included the currently popular student characteristic or characteristics; few studies have developed a theoretical or logical rationale for including a particular set of characteristics. Thus they provide little insight into the processes by which student characteristics affect teacher appraisals or the relative effects of these characteristics. Third, the varying conditions under which the studies have been conducted and the failure to specify the scope of the studies make it difficult to accumulate knowledge on the conditions under which such findings are likely to apply. Fourth, most studies of the influence of student characteristics on teacher appraisals have not considered differences in immediate student performance and behavior in the classroom. Thus it has not been possible to determine if reported differences in teacher appraisals are the result of differences in student characteristics or in actual student performance.

Egan and Archer (1985) observe that the decision to examine teacher appraisals of students using experimental models of prejudice borrowed from social psychology (e.g., Rosenthal and Jacobson, 1968) is in contrast to the study of diagnosis in other professions

where accuracy and rationality of the appraisal are assumed and interest is directed to the strategy of the appraisal process. Egan and Archer (1985) compare teacher appraisals of student ability in mathematics and English with appraisals inferred from standardized tests. They conclude that "...there is little basis for a claim that teachers' ratings are inaccurate--not because their ratings can be shown to be accurate, by reference to some predetermined measure of true ability, but because we cannot produce a rational strategy of classification that is similar to theirs and that gives substantially better results" (Egan and Archer, 1985:32).

Egan and Archer (1985) see little justification for continuing to study teacher ratings of students as a type of irrational cognition. Instead, they suggest research that focuses on the rational aspects of teachers' ratings. For example, in their own study, they observe that teachers were reluctant to use extreme categories and they overused the upper quintiles. Egan and Archer suggest that such patterns might be interpreted in terms of the cognitive psychology of teacher appraisals.

Other studies provide additional examples of how the rational appraisal processes of teachers might be examined. Elmore and Beggs (1972) found that teachers tend to rate students on the most recent incident that reflected a specific behavior rather than on more global behaviors. Natriello and Dornbusch (1983) found that teachers' ratings of students reflected particular classroom behaviors and performance as opposed to general performance and behavior histories

and student status characteristics. Ryan and Levine (1981) studied the impact of sequences of students' past performances on teacher appraisals and found that although the final performance was an important determinant of evaluators' ratings, a simple recency model did not adequately account for all of the data -- prior performance also influenced the appraisal.

Teachers also appear to make important discriminations regarding the quality and nature of the information they use in formulating appraisals. Borko and Shavelson (1978) found that teacher attributions to student ability were influenced by the reliability of the information they had available for assessment. Levin, Imms, and Vilmain (1980) found that college students placed in the teacher role in a series of experiments placed less weight on a source of information seen to be less reliable, but that they did not use the relative variability of scores as an indicator of reliability. Pedulla, Airasian, and Madaus (1980) found that teachers could not separate their judgments about academically related student behaviors observed on a daily basis from their judgments about students' standing on IQ., mathematics, and English, but that teachers could disentangle social behaviors from academically related behaviors.

Studies of the appraisal of student performance might seek to interpret the observed problems in terms of the earlier stages of the evaluation process. For example, Brown (1971) attributes much of the unreliability of teacher appraisals to the fact that teachers use quite different criteria in evaluating students. Starch and

Elliott (1912) relate differences in teacher appraisals to differences in school and teacher standards. Stockhard, Lang, and Wood (1985) found differences in the extent to which student background factors influenced evaluations in English and mathematics, thus suggesting the importance of further study of the role of tasks in the evaluation process. Geisinger and Rabinowitz (1980) found relationships between the type of standards and the sampling method employed by college instructors and the average course grades. Higher grades were given by instructors who adopted criterion-referenced standards (r = .08) or self-referenced standards (r = .28), while lower grades were given by those who adopted norm-referenced standards (r = -.29). Higher grades were given by instructors who sampled student performance through classroom participation (r = .27), term papers and book reports (r= .25), and special projects (r = .37), while lower grades were given by instructors who sampled student performance through examinations and quizzes (r = -.15). Webster and Entwisle (1976), drawing on expectations-states theory (Berger, Cohen, and Zelditch, 1972), develop a theoretical perspective to organize and understand the processes by which appraisals are affected by factors other than objective criteria (e.g., halo and demon effects, Gibb, 1983; Symonds, 1925). Studies of this type provide models of the kind of research that will advance thinking about the appraisal process in the context of an appreciation of the broader evaluation process.

Providing Feedback on Student Performance

An appraisal of a student performance may need to be communicated to various audiences, depending upon the purpose of the evaluation process. Such audiences may include the student, parents, school officials, and potential employers (Ahmann and Glock, 1967). The nature and extent of communications regarding student performance have been the subjects of various investigations and commentaries.

Much of the discussion of feedback on student performance focuses on the visible trappings of traditional evaluation systems in schools and classrooms -- grades and report cards. Jarrett (1963) reviews trends in report cards and notes the movement from reporting based on a percentage system to reporting on the basis of letter grades in secondary schools. He notes the trend in the sixties of moving away from grades toward other methods of reporting. Jarrett (1963) reports on a survey of 258 secondary schools in which it was found that 81% used letters or other symbols, 26% used percentages, 9% used class ranks, 3% used percentile ranks, 2% used written records or logs of student progress, 1% used accomplishment quotients, and 1% used sigma scores. He summarizes then current trends as:

> (1) less frequent reporting for all pupils; (2) more frequent reporting in cases of exceptionally good or exceptionally poor performances; (3) ratings on many more traits and abilities than formerly; (4) making the reports more and more descriptive; and (5) reporting for the purpose of furthering pupil growth (Jarrett, 1963:46).

Chansky (1975) reports on a more recent study of report cards in two percent of school districts nationwide. His analysis considered four major features of reporting: the opening comments, the academic items noted, the personal qualities noted, and the rating systems employed. He found that the use of statements of the purposes of the evaluation declined from the primary grades through high school, the number of academic items marked declined from a high in the primary grades to a low in high school, socio-emotional traits tended to emphasize growth in the lower grades and deviance in the higher grades, and a variety of rating systems were used. In addition, Chansky (1975) classified the rating systems both in terms of the number of categories used and in terms of the content of the category systems. Table 3 presents both the number of rating categories and the content of the categories for the schools in Chansky's survey.

Insert Table 3 About Here

The patterns of responses indicate that the higher the grade level, the more rating categories likely to be used and the greater the variety of reporting systems.

A number of commentators have suggested alternatives to traditional grading and reporting. Rudman (1978) suggests reporting devices such as checklists that are more closely related to the mechanisms for recording student performance. Ediger (1975) suggests more

frequent and more varied mechanisms for reporting student performance, such as telephone and face-to-face conferences with parents. Giannangelo and Lee (1974) and Giannangelo (1975) describe a system of computer-assisted reporting that provides more anecdotal information on student performance. Holtz (1976) presents a reporting method for student performance in element ry science more clearly related to evaluation criteria. Walling (1975) discusses five broad categories of reporting techniques -- traditional grades, percentage ratings, checklists of objectives, narrative evaluations, and conferences. Stewart (1975) describes a multi-dimensional reporting system for use in elementary schools.

Gullickson (1982) reports on the processes used by teachers to provide feedback on tests to students. Most of the teachers in this study provided a grade rather than just a numerical score on tests. In addition, 90% of the teachers reported providing written comments at least occasionally and 55% of the teachers reported providing written comments usually or always. These teachers attempted to provide feedback in a timely manner -- 7% returned tests the same day, 83% returned tests within one day, and only 6% required more than two days to process tests and return them. Gullickson (1982) also asked teachers to classify their feedback activities. The average teacher in his study spent 20 minutes in class review of a test and averaged nine minutes reviewing items selected by the teacher, eight minutes reviewing items questioned by students, and three minutes reviewing the grading procedures. Keep in mind that the teachers in Gullickson's study tended to rely on short answer

and objective tests. In a study of the types of feedback used in classrooms, Zahorick (1968) found that teachers relied upon a limited number of techniques for reviewing test items, and very few teachers indicated why a particular response had merit.

Natriello's (1982) interview study with teachers in four high schools revealed a wide range of activities designed to provide feedback to students. Although most of the teachers used traditional methods to provide feedback (e.g., written comments, conferences, etc.), some had developed innovative techniques. An English teacher provided audio cassette tapes of comments on student papers, and a physical education teacher kept an "open gradebook" that students could examine at any time. Other teachers had students tally their own cumulative scores at various points in the grading period, and still others had students chart their own progress on a regular basis.

A number of observers have remarked on the relationship of the feedback process to other aspects of the evaluation system. Slavin (1978:98) notes that "Feedback is a complex issue, as it has different meanings and uses depending on the way in which it is used." He distinguishes among three kinds of communications regarding student performance as they relate to three purposes of evaluation: "informational feedback," which should tell students where they stand compared to other students and thus should be based on norm-referenced standards; "performance feedback," which should provide students with information on their day-to-day performance and pro-

vide direction for improvement and thus should be based on criterion-referenced standards; and incentive feedback, which should enhance student motivation and thus should be both timely and based on tasks that are neither too difficult nor too easy. Slavin's three types of feedback correspond to the selection, direction, and motivation purposes of evaluation systems noted earlier. Slavin suggests that a single system of evaluation cannot serve all three functions and urges the creation of parallel evaluation systems.

Lissman and Paetziod (1983) also noted the heterogeneous nature of feedback on achievement as it relates to the purposes of evaluation. They distinguished between informative feedback and motivational feedback. Hansen (1977) proposed a system of personalized feedback on achievement that is consistent with the directive purposes of evaluation. Cross and Cross (1980) suggested that teachers who devote more time to writing evaluative comments believe that such feedback will facilitate student motivation.

Relationships between feedback and other aspects of the evaluation process have also been noted. Lintner and Ducette (1974) noted the impact of task variables, particularly task ambiguity, on student responsiveness to praise. Lissman and Paetzoid (1983) observed that certain kinds of feedback seemed more effective for certain kinds of tasks. Oren (1983:307) noted the relationship between "rich, more specific, and individualized" feedback and the motivational purposes of evaluation, specifically, the attributional tendencies of students.

The Effects of Evaluation Processes on Students

Although the evaluations that take place in schools and classrooms clearly have powerful effects on students and others (e.g., see Poole, 1979), consideration of studies of these effects has been deferred until now for several reasons.

First, relatively little descriptive information on evaluation processes in schools and classrooms has been considered in designing effects studies, even though many studies seek to create new knowledge as the basis for improved practice. Thus, the descriptive information on evaluation in schools and classrooms reviewed above provides important groundwork for consideration of the effects studies. For example, some studies seek to develop alternatives to norm-referenced standards, but descriptive accounts suggest that such standards may not now be used extensively by teachers.

Second, most of the effects studies concentrate on only one or two aspects of the evaluation process outlined above, and thus fail to consider the impact of other key elements. The conclusions drawn from such studies should be approached with caution. For example, few studies consider the nature of the assigned tasks upon which students are being evaluated, yet it is clear that task differences condition the impact of evaluation processes.

Third, few of the effects studies consider the multiple purposes of evaluations in schools and classrooms. They often compare the impact of different evaluation methods on some outcome that has

nothing to do with the purpose for which one of the methods was developed. For example, a study demonstrating that differentiated feedback contributes more to directing future student performance than a single letter grade may be simply showing that an evaluation system created for the purpose of providing direction to students does a better job of providing that direction than another evaluation system created for the purpose of selecting students.

With the above reservations clearly in mind, it is useful to review the effects of some selected aspects of evaluation processes.

Investigators are only beginning to recognize the importance of classroom tasks in understanding educational and evaluation processes (Doyle, 1983). A particularly interesting line of research in this area focuses on the impact of the task structure of classrooms on students' conceptions of the distribution of ability in the class. In a study of fifth- and sixth-graders, Rosenholtz and Wilson (1980) found that in classes characterized by what they called higher "resolution" (i.e. less task differentiation, more ability grouping, more evaluations comparing the work of one student with another, and less student autonomy to choose tasks) there was higher concurrence among classmates, between self and classmates, between teacher and classmates, and between self and teacher in ratings of reading ability. Rosenholtz and Rosenholtz (1981) found that these same high "resolution" classroom structures led to more dispersed evaluations of reading ability by students themselves, by classmates, and by teachers. They also found that low "resolution"

(dimensional) classroom structures diminished the effect of evaluations by the teacher on peer evaluations of an individual's reading ability.

In a study of third grade classrooms Simpson (1981:127) found that low levels of curricular differentiation (one element of unidimensional classroom structure) led to "...a more nearly normal distribution of self-reports of ability by increasing the proportion of students reporting ability levels below average and far below average." Moreover, low curricular differentiation also appeared to lead to a more generalized view of academic ability, greater peer consensus about students' performance levels, and to greater influence of peers on individual's self-reported ability. These studies suggest that the consistency or differentiation of task assignments, criteria, standards, sampling strategies, and feedback mechanisms may affect the perceived distribution of ability.

Dornbusch and Scott (1975) make the point that criteria add to the definition of the assigned task and direct the attention of performers to the key elements of the task for which they will be held accountable. Schunk (1983) reports on a study in which some children were offered rewards for participating in a task, others were offered rewards for careful work on the task, and still others were not offered rewards until they had completed the task. The results indicated that the first group of children, who had received both a task assignment and information on the criteria for performance, showed the highest levels of skill, self-efficacy, and rapid problem solving.

This should not be surprising. As Deutsch (1979:396) points out, "students are in a bewildering position if a teacher marks them without telling them in sufficient detail the values, rules, and procedures employed in his or her grading. In such a situation, the mark-oriented students are necessarily anxiously dependent on the teacher's approval, since they have no other basis for guiding their behavior to achieve merit... Where the instructor is explicit in his or her style of grading, the student can be more independent of the teacher."

Natriello (1982) found that more than 30% of the students in his study of four suburban high schools reported that they had received unsatisfactory evaluations because they had misunderstood the criteria by which they were to be evaluated. Smith (1984) observed that clarity has been demonstrated to be an important component of teaching in research on teaching effectiveness (Rosenshine and Furst, 1971). Smith studied the impact of teacher "use of uncertainty phrases" on student achievement and found that such phrases negatively affected achievement.

However, explicitness may have undesirable effects as well. Deutsch (1979) notes that explicit evaluation systems may lead mark-oriented students to limit their work to what is being assessed by the procedures employed in the grading or to attempt to outwit the procedures. He cites as an example managers in the Bell System who are graded or evaluated by "profit indices" and who often outwit the system by postponing routine maintenance costs, which results in

equipment breakdowns several years later when successful managers have moved on to new positions. Deutsch (1979) concludes that such dilemmas are avoidable only to the extent that the evaluation system fosters the motivation to achieve intrinsic merit rather than its external symbols.

The effects of performance standards seem to be more complex than is typically thought. Investigations have focused on both the level of standards and the type of standards used in evaluation systems. Early studies of the impact of school standards on student performance (Brookover and Schneider, 1975) seem to have survived the challenge that the correlation between teacher standards and student performance could result from the impact of the latter on the former (Crano and Mellon, 1978). Findings from the school effectiveness literature (Purkey and Smith, 1983), the teacher expectations literature (Brophy and Evertson, 1981) and the task goals literature (Locke, 1968; Rosswork, 1977) suggest that higher standards yield better student performance. In studies specifically focused on evaluation processes, Natriello and Dornbusch (1984) found that higher standards led to greater student effort on school tasks and to students being more likely to attend class, and Natriello and McDill (1986) found that when teachers had standards for homework, students were more likely to spend time on homework.

However, the effects of higher standards may not be uniformly positive. Natriello (1982) found that students who perceived standards for their performance as unattainable were more likely to

become disengaged from high school. McDill, Natriello, and Pallas (1985) suggested that higher standards may actually have detrimental effects for at-risk students in secondary schools. There seems to be a curvilinear relationship between the level of standards and student effort and performance. The goal would seem to be to challenge students without frustrating them (Atkinson, 1958).

The impact of different types of standards has also been investigated. Perhaps the most attention has been devoted to norm-referenced standards or "grading on the curve." Michaels (1977) designates the reward structure associated with this practice as "individual competition, in which grades are assigned to students based on their performances relative to those of their classmates" and distinguishes it from "individual reward contingencies, in which grades are assigned to students on the basis of how much material each student apparently masters." He considers the effects of these two reward structures along with two other reward structures (group competition and group reward contingencies) on student academic performance. Reviewing the relevant literature, he concludes that individual competition consistently produces superior academic performance. However, he observes that the superior academic performance found to be associated with individual competition may be limited to the top third of the class, to those students who are most responsive to the reward structure, for several reasons: First, the value of grades may vary considerably across students; second, the probability of receiving high grades also varies considerably across students; third, performance gains by initially low-performing stu-

dents may be seldom reinforced in systems of individual competition. Michaels concludes by arguing that the reward structure itself may be less important than seeing to it that the rewards selected are valued by all students, are made contingent on the performance to be strengthened, and that significant performance gains are intermittently reinforced.

Deutsch (1979) criticizes individual competition or grading on the curve as an artificially created shortgage of good grades. He argues that the "Disappointing rewards, induced by an artificial scarcity, are likely to hamper the development of educational merit and the sense of one's own value." (Deutsch, 1979:394). Moreover, under individual competition, "students are more anxious, they think less well of themselves and of their work they have less favorable attitudes toward their classmates and less friendly relations with them, and they feel less of a sense of responsibility toward them." (Deutsch, 1979:399)

Examining the same studies as Michaels, Duetsch (1979:398) concludes that a number of these studies were flawed because they did not equate the objective probability of reward in the reward structures being compared. Deutsch's reanalysis of these studies shows "no systematic differences in performance on isolated work under several different reward systems." Williams, Pollack, and Ferguson (1975) also found no significant differences between the achievement and self-reported attitudes or school-related behavior of students exposed to norm-referenced and criterion-referenced standards. They

also found that criterion-referenced standards enabled some students who performed poorly initially to increase their performance on later tests, but students who did well initially began to work less hard than students working under a norm-referenced system who had to deal with the possibility that other students would try harder on the next test and raise the curve.

Deutsch (1979:394) also argues that the competitive struggle for scare goods in the classroom teaches students about more than just their own performance. He notes that they "...are socialized into believing that this is not only the just way but also the natural and inevitable way of allocating scarce values in the larger, impersonal, nonfamiliar world. They also learn that there are winners and losers in such competitions and that, although it is possible for them to win, they are more likely to lose."

Finally, Deutsch (1979) points out that the artificially induced scarcity of grades lends them importance. In fact, it is one of the chief means of conveying meaning to grades, which themselves are typically of uncertain quality and unspecific meaning.

Norm-referenced standards have also been compared to individually-referenced standards for their effects on student performance. Beady, Slavin, and Fennessey (1981) found no differences in the effects of norm-referenced standards and individually-referenced standards among students participating in a program of focused instruction, a particular model of direct instruction. On the other hand, under different task conditions Rheinberg (1983) found that

students working under individually-referenced standards showed more realistic strategies of goal setting, more often attributed their successes to their own effort, and performed better than students working under norm-referenced standards.

Bolocofsky and Mescher (1984) added complications to the issue of the impact of different standards by considering the effects of different standards for students who differ in self-esteem and locus of control. They found that students with different characteristics performed differently under different kinds of standards. Self-referenced standards worked best with students with low self-esteem and internal locus of control. Criterion-referenced or absolute standards worked best with students with low self-esteem and external locus of control. Norm-referenced standards worked best with students with high self-esteem regardless of locus of control.

Many studies have examined the impact of different types of standards on student cooperation and competition. These studies typically examine the relationships between the evaluations made and rewards distributed and the tendency for students to perform tasks independently, cooperatively, or competively. Slavin (1977:634) in a review of much of this research uses the term "interpersonal reward structure" to refer to the dependence or lack of dependence of any given student on any other student. He distinguishes three types of interpersonal reward structures: competitive reward structures, where the probability of one student receiving a reward is negatively related to the probability of other students receiving a

reward; independent reward structures, where the probability of one student receiving a reward is unrelated to the probability of other students receiving a reward; and cooperative reward structures, where the probability of one student receiving a reward is positively related to the probability of other students receiving a reward.

Slavin (1977:644) reviewed the research on the impact of these reward structures on student social behavior and academic performamce in the classroom. He concluded that cooperative structures enhance social behavior along a number of dimensions, including interpersonal attraction, friendliness, positive group evaluation, helpfulness, and cross-racial interaction. Competitive and independent reward structures were found to be more effective in increasing performance when tasks required little cooperation or when there was little opportunity to share resources to facilitate performance, but Slavin noted that cooperative structures should be effective in promoting performance when such cooperation and sharing are necessary and permitted.

A number of investigations have focused on the frequency of the sampling process, especially the frequency of testing. Reviewers of the research on the frequency of testing (Feldhusen, 1964; Peckham and Roe, 1977) have found that although early studies of testing frequency indicated that more frequent testing had uniformly positive effects on student learning and motivation, more recent studies incorporating more variables suggest that more frequent testing may

not benefit all students in all contexts. However, considering evaluation activities as contests, Deutsch (1979:396) concludes that "The existence of many diverse contests diffuses competition and reduces the negative implications of any particular contest: It is less harmful to one's self-esteem and social standing."

Studies of testing frequency have not typically viewed testing as part of a larger evaluation process. In the model developed here, however, testing is merely one method of sampling student performance and outcomes. Viewed in this way, the frequency of testing issue can be more appropriately stated as one of selecting an appropriate interval to collect samples of student performance on particular tasks to be evaluated in terms of particular criteria. Certain student tasks may require more extensive and/or more frequent sampling procedures to insure that the appraisal process is based on valid and reliable samples of student performances and outcomes. Objections that frequent evaluation raises student anxiety must be balanced against the preferences of students that the teacher have more extensive and more representative samples of their work. Of course, overly frequent evaluation may have negative effects on student motivation and performance when it disrupts performance itself.

Consideration of the <u>appraisal</u> process focuses attention on the connection between student performances and the evaluations made of those performances by teachers, often from the perspective of the teacher attempting to carefully relate performance information to predetermined tasks, criteria, and standards. The quality of the

connection between student performance and evaluations also appears to have important effects on students. Natriello and Dornbusch (1984) found that when students perceived the evaluations of their performance on school tasks to be unsound (i.e., not to accurately reflect their effort and performance), they were less likely to consider these evaluations important and less likely to devote effort to the associated tasks.

But these effects may be more complicated as indicated by work on the theory of learned helplessness which suggests that experiencing uncontrollable outcomes should depress performance (Abramson, Seligman and Teasdale, 1978, Seligman, 1975), and by work which suggests that experiencing uncontrollable outcomes facilicates increased performance by producing an increased need for control (Roth and Bootzin, 1979; Thornton and Jacobs, 1972). An integrative model developed by Wortman and Brehm (1975) suggests that brief exposure to uncontrollable outcomes will lead to improved performance while extended exposure will lead to decreased performance. Research involving high school students (Buys and Winefield, 1982) finds only decreased student performance in reaction to the experience of uncontrollable outcomes, a pattern the authors link to the relatively less self-reliant and less self-confident nature of high school students compared to adults, and to the nature of the school environment, which they see as tending to foster helplessness.

Students may differ in their perceptions of appraisal processes independent of the process itself. Evans and Engelberg (1985) found

that older and higher-achieving students understood grading practices better than younger and lower-achieving students, and that younger and lower-achieving students were more likely to attribute grades to external and uncontrollable factors while high achievers and older students attributed grades to internal and controllable factors.

A number of studies have examined the impact of the feedback presented as part of the evaluation process. Stewart and White (1976) present the results of their own study and review those of 12 others which attempted to replicate Page's (1958) classic study of the effects of feedback. Page found that "When the average secondary teacher takes the time and trouble to write comments (believed to be "encouraging") on student papers, these apparently have a measurable and potent effect upon student effort, or attention, or attitude, or whatever it is which causes learning to improve..." (Page, 1958:180-181). Stewart and White (1976) reach a slightly less confident conclusion, noting that the positive effect obtained by Page may depend upon the particular learning conditions and the nature of the teacher comments. Cross and Cross (1980) found that personalized encouraging comments from the teacher used in addition to a grade on tests and assignments enhanced the "internality" of students in an inner-city junior high school.

Feedback may also affect students in schools and classrooms other than those to whom the feedback pertains. A study of third graders by Simpson (1981) illustrates how evaluative feedback decisions can

affect students' perceptions of the ability levels of their classmates. Simpson (1981:124) argues that "Grades are singular symbols taking on unidimensional comparative meaning from the abstract numerical system which defines them. Frequent grading is capable of reducing even relatively complex performances to a single dimension, because grades reduce information to numbers, because these numbers can be averaged, and because teachers and student peers can use these numbers to place students on a single global stratification scale." Simpson finds that in classrooms where teachers report "always" or "usually" grading student work (as opposed to those in which they "never" or "seldom" grade such work), where they report using few kinds of instructional materials, and where they seldom use alternative media and seldom allow students to choose their tasks, there was greater dispersion among students' reported ability levels, greater generalization of students' reported ability levels, greater peer consensus as to students' relative performance levels, and greater peer influence over students' reported ability levels. Thus, the use of grades seems to lead to more pronounced and more powerful ability stratification processes in the classroom.

A similar effect on the distribution of attributional tendencies in classrooms was found by Oren (1983), who explored the effects of evaluation feedback on the attributional tendencies of students. Results indicated that in classrooms with differentiated, specific, and individualized feedback, the attributional tendencies of low achievers were more like those of high achievers. Specifically, low achievers in such classrooms scored higher on internal control than

did low achievers in classrooms with less differentiated feedback systems.

The affective value of feedback has also been shown to affect attributions in classrooms. Meyer, Bachmann, Biermann, Hempelmann, Ploger, and Spiller (1979) report on a series of six experimental studies which investigated the extent to which praise and criticism in response to task performance provided information about other's perceptions of a focal actor's ability. In these studies subjects were presented with descriptions of two students who had obtained identical results at a task. One of the students received neutral feedback while the other was praised for success or criticized for failure. Studies using adult subjects revealed that praise after success and neutral feedback after failure led to the perception that the focal actor's ability was low, and neutral feedback for success and criticism after failure led to the perception that the focal actor's ability was high. However, these findings varied by the age of the respondents. Third-grade students believed that the student praised by the teacher was the brighter one; students in grades 4 to 7 selected the praised student and the student receiving neutral feedback in approximately equal numbers; and students in grades 8 and above believed that the student receiving neutral feedback was brighter than the one receiving positive feedback following successful performance.

Although the effects of feedback in the classroom appear to be powerful, they are multidimensional and complex. Simple injunctions

to increase feedback for one purpose or another are likely to set in motion a range of processes that need further examination.

Although the above studies of the effects of aspects of the evaluation process have suggested some possible consequences for certain evaluation processes, little attention has been devoted to developing an understanding of entire evaluation systems composed of purposes, tasks, criteria, standards, samples, appraisals, and feedback. One of the key issues to be examined in thinking about systems of evaluation is the relationship between various aspects of the process and the extent to which there is consistency among them. For instance, evaluations and evaluation systems may differ in consistency between task assignments and criteria set for the task. Some teachers may take care that the performance criteria set for a task be appropriate to the nature of the task assignment, while others may not -- a teacher may designate a task as a creative opportunity when an assignment is made but hold students accountable for a formulaic set of criteria. A second instance might be the consistency between the criteria and standards set for the task and the process of sampling student performances and outcomes. A teacher may specify criteria related to the actual performance of the task (e.g. how to proceed to solve a math problem), but only sample the outcome of the performance (e.g. the correctness of the answer).

Little research has examined the extent to which teachers implement a consistent system of performance evaluation for students. Interviews conducted by Natriello (1982) with 80 secondary school

teachers suggest that teachers vary widely in their ability to articulate a systematic approach to the evaluation of student performance. Also, examinations of teacher preparation curricula indicate that prospective teachers receive little or no training in the evaluation of student performance (Mayo, 1967; Roeder, 1973). The effects of this lack of consistency could be quite negative. Natriello (1982) reported that high school students who experienced more inconsistencies in the evaluation system were also more likely to become disengaged from school. In that study students were asked to report on the extent to which they perceived incompatibilities or inconsistencies in the evaluation processes to which they were subjected. Students who reported being exposed to such incompatibilities were more likely to report complaining to other students about the evaluation and authority system of the school.

The potential consequences of inconsistencies in systems of evaluation and the likelihood that such inconsistencies are widespread make it particularly important to consider evidence on how different components of evaluation systems might fit together to produce a coherent evaluation process. The best evidence of such systems comes from formal programs and policies rather than from studies of particular elements of evaluation processes.

The Impact of Programs and Policies

Even though major educational programs and policies seldom have an explicit focus on evaluation, consideration of programs and policies that might affect evaluation processes in schools and classrooms provides a perspective different from those studies of evaluation processes reviewed thus far. These comprehensive programs typically address (at least implicitly) multiple elements of the evaluation process as opposed to individual features.

Most major programs include a rationale which involves some statement of purpose. Many programs entail a conception of the nature of academic tasks in schools and classrooms, a particular type of standard for performance, and guidelines for the type of feedback that students should receive. Several major programs can be considered for their effects on the evaluation practices of educators and ultimately on student learning both to illustrate the utility of the model of evaluation processes specified earlier and to reveal more about the implications of the programs for evaluation processes.

Table 4 presents a summary of the implications of three major programs or policies for evaluation processes in schools and classrooms.

Insert Table 4 About Here

Minimum competency testing programs are enacted for the purposes of certification of students. They tend to involve relatively simple tasks with time limits on performance and absolute standards. Such programs are based on infrequent samples of performance, rely on appraisals prepared by individuals other than the immediate teacher of the subject, and utilize simple feedback to students.

Mastery learning programs are implemented for the purpose of providing students with direction. They tend to structure the cu lum in terms of relatively small discrete tasks with criteria that do not involve time limits on performance but do inv ve absolute standards. Mastery learning programs are based on quite frequent samples of performance, provide "A's" for all students who master the material, and utilize frequent and differentiated feedback to orient students to their accomplishments and remaining needs.

Public Law 94-142 was enacted to require individualized instruction and evaluation for handicapped students in the least restrictive environment by providing greater direction to such students. The policy implies individualized tasks with non-specified criteria and individually referenced standards. Further, P.L. 94-142 envisions frequent sampling of student performance and frequent feedback to students of the appraisals of their individual teachers.

This brief analysis of the evaluation implications of these three major educational programs or policies reveals several advantages of the application of the evaluation framework. First, examining programs in terms of the elements of the evaluation allows for a clear

specification of the purposes of different programs so that programs with different purposes are not as likely to be examined for effects they are not designed to have. For example, Table 4 makes it clear that a study comparing a mastery learning program to a minimum competency testing program could not fairly look for the same effects from both programs.

Second, considering programs and policies in terms of elements of the evaluation process allows identification of areas in which the programs and policies carry few implications for evaluation systems and thus areas where differences in practice may lead to quite different outcomes from the same program. For example, none of the three programs in Table 4 carry very specific implications for evaluation criteria. As a result individual implementations of these programs might vary considerably and produce quite different outcomes from what are ostensibly the same type of programs.

Third, examining programs in terms of the elements of the evaluation process facilitates the identification of conflicts between different programs when they are implemented simultaneously. For example, in some states teachers are simultaneously subjected to the requirements of minimal competency testing programs and P.L. 94-142. The former attempts to implement absolute standards; the latter mandates individually referenced standards. Teachers are likely to experience considerable conflect trying to comply with both programs (Bender, 1984).

Overall, the evaluation framework provides one way to link general educational programs and policies to the specific practices of local educators in schools and classrooms. Analyzing newly proposed educational programs and policies for their implications for classroom evaluation processes should reveal much about the problems of implementation as well as about the likely effects of such programs on students.

Conclusions

Evaluation processes in schools and classrooms are both complex in their organization and powerful in their effects on the lives of educators and students. This review demonstrates that despite examination of elements of evaluation systems by practitioners and researchers, a comprehensive and powerful conceptual framework to facilitate the study of student evaluation practices and their effects has not yet been developed. The framework described here is a first step in that direction. Further refinement and elaboration may permit more keenly drawn conclusions about evaluation processes. Students, teachers, and administrators will continue to encounter the influences of evaluation processes as they work in schools and classrooms. Educational researchers will continue to have their studies affected by various evaluation practices. The only question is whether practitioners or researchers will mount the effort to secure greater understanding of and control over the evaluation processes that affect us all.

FOOTNOTES

<1> Dornbusch and Scott (1975) note that the term "task conceptions" represents a compromise between the notion of a task as completely objective and the notion of the understanding of a task as completely subjective.

<2> This overly empiricist approach is merely a specific manifestation of a more general phenomenon identified by Lakatos (1971).

<3> The idea that there is a danger in allowing the match between instructional materials and test items (Linn, 1983:187) could only arise in situations in which there is a failure to develop genuine criteria separate from objectives rooted in measurable behavior. Such conditions arise because of the overly empiricist approach in which there is a one-to-one correspondence between criteria and indicators.

<4> The term "criterion-referenced testing" combines the word "criterion," the concept of a standard, and a technique of sampling all in a single phrase.

REFERENCES

Abramson, L.Y., Seligman, M.E.P., & Teasdale, J.D. (1978). Learned helplessness in humans: Critique and reformulation. Journal of Abnormal Psychology, 87, 49-74.

Ahmann, J., & Glock, M.D. (1967). Evaluating pupil growth: Principles of tests and measurements (fourth edition). Boston: Allyn and Bacon.

Airasian, P.W., & Madaus, G.F. (1983). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20, 103-118.

Allington, R.L. (1978). Teacher ability in recording oral reading performance. Academic Therapy, 14, 187-192.

Armbruster, B.B., Stevens, R.J., & Rosenshine, B. (1977). Analysing content coverage and emphasis: A study of three curricula and two tests. Technical Report Number 26. Urbana: Center for the Study of Reading, University of Illinois.

Arnold, M. (1983). Statistical models of fairness and their impact on non-biased assessment. Diagnostique, 8, 150-158.

Atkinson, J.W. (1958). Towards experimental analysis of human motivation in terms of motives, expectancies, and incentives. In J.W. Atkinson (ed.), Motives in fantasy, action and society. Princeton, NJ: Van Nostrand..

Barnhard, J.W., P.G. Zimbardo, & Sarason, S.B. (1968). Teachers' ratings of student personality traits that relate to IQ and social desirability. Journal of Educational Psychology, 59, 128-132.

Beady, C.J., Slavin, R.E., & Fennessey, G.M. (1981). Alternative student evaluation structures and a focused schedule of instruction in an inner-city junior high school. Journal of Educational Psychology, 75, 518-523.

Beady, C., & Slavin, R. (1981). Making success available to all students in desegregated schools: An experiment. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Beck, M.D., & Stets, F.P. (1979). Teacher opinions of standardized test use and usefulness. Paper presented at the annual meeting of the American Educational Research Association, San Fransico, April.

Beggs, D.G., Mayer, R., & Lewis, E.L. (1972). The effects of various techniques of interpreting test results on teacher perception and pupil achievement. Measurement and Evaluation in Guidance, 5, 290-297.

Berger, J., Cohen, B.P., & Zelditch, M. (1972). Status characteristics and social interaction. American Sociological Review, 37, 241-255.

Berk, R.A. (1976). Determination of optional cutting scores in criterion-referenced measurement. Journal of Experimental Education, 45, 4-9.

Bidwell, C.E. (1965). The school as a formal organization. Pp. 972-1022 in J.G. March (ed.), Handbook of organizations. Chicago: Rand McNally.

Bloom, B.S. (1968). Learning for mastery. Evaluation Comment, 1, 2.

Bloom, B.S. The new direction in educational research: alterable variables. Phi Delta Kappan, 61, 382-385.

Bolocfsky, D.N., & Mescher, S. (1984). Student characteristics: using student characteristics to develop effective grading practices. The Directive Teacher, 6, 11-23.

Borko, H., & Shavelson, R.J. (1978). Teachers' sensitivity to the reliability of information in making causal attributions in an achievement situation. Journal of Educational Psychology, 70, 271-279.

Bresee, C.W. (1976). On "Grading on the Curve." The Clearing House, 5, 108-110.

Brookover, W.B., Schneider, J.M. (1975). Academic environments and elementary school achievement. Journal of Research and Development in Education, 9, 82-91.

Brophy, J., & Evertson, C. (1981). Student Characteristics and Teaching. New York: Longman.

Brown, D.J. (1971). Appraisal Procedures in the Secondary Schools. Englewood Cliffs, NJ: Prentice-Hall.

Brown, A., & Craig, R.P. (1977). Grading testing and grading. Elementary School Journal, 77, 395-399.

Burton, N.W. (1978). Societal standards. Journal of Educational Measurement, 15, 263-271.

Buys, N., & Winfield, A.H. (1982). Learned helplessness in high school students following experience of noncontingent rewards. Journal of Research in Personality, 6, 6-9.

Chansky, N.M. (1975). A critical examination of school report cards from K through 12. Reading Improvement, 12, 184-192.

Crano, W.D., & Mellon, P.M. (1978). Causal influence of teachers' expectations on children's academic performance: A cross-lagged panel analysis. Journal of Educational Psychology, 70, 39-49.

Crooks, A.D. (1933). Marks and marking systems: A digest. Journal of Educational Research, 27, 27:259-272.

Cross, L.J., & Cross, G.M. (1980). Teachers' evaluative comments and pupil perception of control. Journal of Experimental Education, 49, 68-71.

Culler, J. Literary competence. Pp. 101-117 in J.P. Thompkins (ed.), Reader-response criticism: From formalism to post structuralism. Baltimore, MD: Johns Hopkins University Press. 101-117.

Davis, R.B., & McKnight, C. (1976). Conceptual, heuristic, and S-algorithmic approaches in mathematics teaching. Journal of Children's Mathematical Behavior, 1(Supplement 1), 271-286.

Deutsch, M. (1979). Education and distributive justice: Some reflections on grading systems. American Psychologist, 34, 391-401.

Dornbusch, S.M., & Scott, W.R. (1975). Evaluation and the exercise of authority. San Francisco: Jossey-Bass.

Doyle, W. (1983). Academic work. Review of Educational Research, 53, 159-199.

Doyle, W. (1979). The tasks of teaching and learning in classrooms. Research and Development Report Number 4103. Austin, Texas: Research and Development Center for Teacher Education, University of Texas, 1979.

Ediger, M. (1975). Reporting pupil progress: Alternatives to grading. Educational Leadership, 32, 265-267.

Egan, O., & Archer, P. (1985). The accuracy of teachers' ratings of ability: A regression model. American Educational Research Journal, 22, 25-34.

Elmore, P., & Beggs, D.L. (1972). Stability of teacher ratings of pupil behavior in a classroom setting. Paper presented at the meetings of the American Personnel and Guidance Association, March.

Evans, E.D., & Engelberg, R.A. (1985). A developmental study of student perceptions of school grading. Paper presented at the Biennial Meeting of the Society for Research on Child Development. Toronto.

Feldhusen, J.F (1964). Student perceptions of frequent quizzes and post-mortem discussions of tests. Journal of Educational Measurement, 1, 51-54.

Fennessey, J. (1973). The "Focused Flexibility and GREG" project at Walbrook High School. Summary Report. Baltimore, Maryland: Center for the Social Organization of Schools.

Feshbach, N.D. (1969). Student teacher preferences for elementary school pupils varying in personality characteristics. Journal of Educational Psychology, 60, 126-132.

Frase, L.T. (1972). Maintenance and control in the acquisition of knowledge from written materials. In J.B. Carroll, and R.O. Freedle (eds.), Language Comprehension and the Acquisition of Knowledge. Washington, DC: Winston.

Frase, L.T. (1975). Prose processing. Pp. 1-48 in G.H. Bower (ed.), The Psychology of Learning and Motivation, Volume 9. New York: Academic Press. 1-48.

Frederiksen, C.H., & Dominic, J. (1981). Writing: The Nature, Development, and Teaching of Written Communication (volume 2). Hillsdale, NJ: Erlbaum.

Fuchs, L.S., Deno, L., & Mirkin, P.K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. American Educational Research Journal, 21, 449-460.

Gaston, N. (1976). Evaluation in the affective domain. Journal of Business Education, 52, 134-136.

Geisinger, K.F., & Rabinowitz, W. (1980). Individual differences among college faculty in grading. Journal of Instructional Psychology, 7, 20-27.

Giannangelo, D.M. (1975). Make report cards meaningful. The Educational Forum, 39, 409-415.

Giannangelo, D.M., & Lee, K.Y. (1974). At last: Meaningful report cards. Phi Delta Kappan, 55, 630-631.

Gibb, G.D. (1983). Influence of "halo" and "demon" effects in subjective grading. Perceptual and Motor Skills, 56, 67-70.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. American Psychologist, 18, 519-521.

Glass, G.V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Gronlund, N.E. (1971). Measurement and Evaluation in Teaching. (second edition). New York: Macmillan.

Gullickson, A.R. (1982). The practice of testing in elementary and secondary schools. Unpublished Report. ED229391.

Gullickson, A.R. (1984). Teacher perspectives of their instructional use of tests. Journal of Educational Research, 77, 244-248.

Hackman, J.R. (1969). Toward understanding the role of tasks in behavioral research. Acta Psychologica, 31, 97-128.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48, 1-47.

Hansen, J.M. (1977). Personalized achievement reporting: Grades that are significant. The High School Journal, 60, 255-263.

Heller, L. (1978). Assessing the process and the product: An alternative to grading. English Journal, 6, 66-69.

Hempel, C.G. (1952). Fundamentals of Concept Formation in Empirical Science. Chicago: University of Chicago Press.

Herman, J., & Dorr-Bremme, D.W. (1984). Testing and assessment in American public schools: Current practices and directions for improvement. Los Angeles: Center for the Study of Evaluation, University of California at Los Angeles.

Holmes, M. (1978). Evaluating students in the affective domain. School Guidance Worker, 33, 50-58.

Holtz, R.E. (1976). More than a letter grade. Social Education, 14, 23-24.

Jackson, G.B. (1975). The research evidence on the effects of grade retention. Review of Educational Research, 45, 613-635.

Jarrett, C.D. (1962). Marking and reporting practices in the American secondary school. Peabody Journal of Education, 41, 36-48.

Jenkins, J.J. (1977). Remember that old theory of memory? Well, forget it! Pp. 413-430 in R. Shaw and J. Bransford (eds.), Perceiving, acting, and knowing: toward an ecological psychology. Hillsdale, N.J.: Erlbaum, 1977. 413-430.

Johnson, J.R. (1984). Synthesis of research on grade retention and social promotion. Educational Leadership, 41, 66-68.

Krathwohl, D.R., & Payne, D.A. (1971). Defining and assessing educational objectives. p. 17-45 in Robert L. Thorndike (ed.), Educational Measurement (second edition). Washington, DC: American Council on Education.

Ladd, E. (1961). A comparison of two types of training with reference to developing skill in diagnostic oral reading testing. Unpublished doctoral dissertation, Florida State University, 1961.

Lakatos, I. (1971). Pp. 91-196 in I. Lakatos and A. Musgrave (eds.), Criticism and the growth of knowledge. New York: Cambridge U. Press.

Lawler, E.E. (1976). Control systems in organizations. Pp. 1247-1291 in M.D. Dunnette (ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally.

Leinhardt, G., & Seewald, A.M. (1981). Overlap: What's tested, what's taught? Journal of Educational Measurement, 18, 85-96.

Levin, I.P., Ims, J.R., & Vilmain, J.A. (1980). Information variability and reliability effects in evaluating student performance. Journal of Educational Psychology, 72, 355-361.

Levine, M. (1976). The academic achievement test: Its historical context and social functions. American Psychologist, 31, 228-238.

Levine, A., & Levine, M. (eds.). (1970). The Gary Schools. Epilogue by Abraham Flexner and Frank P. Bachman. Cambridge, MA: MIT Press.

Lien, A.J. (1967). Measurement and evaluation of learning: A handbook for teachers. Dubuque, IA: Wm. C. Brown.

Lindquist, E.F. (1969). The impact of machines on educational measurement. Pp. 351-369 in R.W. Tyler (ed.), Educational evaluation: New roles, new means. The 68th Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press. 351-369.

Lindvall, C.M. (1961). Testing and evaluation: An introduction. NY: Harcourt, Brace, and World.

Linn, R.L. (1983). Testing and instruction: Links and distinctions. Journal of Educational Measurement, 20, 179-189.

Lintner, A.C., & Ducette, J. (1974). The effects of locus of control, academic failure and task dimensions on a student's responsiveness to praise. American Educational Research Journal, 11, 231-239.

Lissman, U., & Paetzoid, B. (1983). Achievement feedback and its effects on pupils -- a quasi-experimental and longitudinal study of two kinds of differential feedback, norm-referenced and criterion-referenced feedback. Studies in Educational Evaluation, 9, 209-222.

Locke, E.A. (1968). Toward a theory of task motivation and incentives. Organizational Behavior and Human Performance, 3, 157-189.

Mager, R.F. (1962). Preparing instructional objectives. Palo Alto, CA: Feardon.

Mayo, S.T. (December 1967). Pre-service preparation of teachers in educational measurement. Chicago: Loyola University.

McDill, E.L., Natriello, G., & Pallas, A. (1985). Raising standards and retaining students. Review of Educational Research, 55, 415-434.

Meskauskas, J.A. (1976). Evaluation models for criterion-referenced testing: views regarding mastery and standard setting. Review of Educational Research, 46, 133-158.

Meyer, W., Bachmann, M., Biermann, U., Hempelmann, M., Ploger, F., & Spiller, H. (1979). The informational value of evaluative behavior: Influences of praise and blame on perceptions of ability. Journal of Educational Psychology, 71, 259-268.

Michaels, J.W. (1977). Classroom reward structures and academic performance. Review of Educational Research, 47, 87-98.

Morine-Dershimer, G. (1979). How teachers see their pupils. Educational Research Quarterly, 3, 43-53.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: U.S. Government Printing Office.

National Institute of Education. (1979). Testing, teaching, and learning. Report of a Conference on Research on Testing. Washington, DC: National Institute of Education.

Natriello, G. (1985). Merit pay for teachers: The implications of theory for practice. In H.C. Johnson, Jr., (ed.), Merit, money and teachers' careers. Sanham, MD: University Press of America.

Natriello, G. (1982). Organizational evaluation systems and student disengagement in secondary schools. St. Louis, MO: Washington University, Final Report to the National Institute of Education.

Natriello, G., & Dornbusch, S.M. (1984). Teacher evaluative standards and student effort. NY: Longman.

Natriello, G., & McDill, E.L. (1986). Performance standards, student effort on homework and academic achievement. Sociology of Education, 59, 18-31.

Natriello, G., & McPartland, J. (1987). Adjustments in High School Teachers' Grading Criteria: Accomodation or Motivation? Paper presented at the Annual Meeting of the American Educational Research Association. Washington, D.C.: April.

Nearine, R.J. (1970). The test, the time, and the teacher. Measurement and Evaluation in Guidance, 2, 214-216.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Olejnik, S.F. (April 1979). Standardized achievement programs viewed from the perspective of non-measurement specialist. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Openshaw, K. (1967). A failure of the Minnesota Teacher Attitude Inventory to relate to teacher behavior. Journal of Teacher Education, 18, 233-239.

O'Regan, M., Airasian, P., & Madaus, G. (April 1979). The use of standardized test information by Irish teachers. Paper presented at the annual meeting of the National Council on Measurement and Education, San Francisco.

Oren, D.L. (1983). Evaluation systems and attributional tendencies in the classroom: A sociological approach. Journal of Educational Research, 76, 307-312.

Page, E.B. (1958). Teacher comments and student performance: A seventy-four classroom experiment in school motivation. Journal of Educational Psychology, 49, 173-181.

Page, W, & Carlson, K. (1975). The process of observing oral reading scores. Reading Horizons, 15, 147-150.

Peckham, P.D., & Roe, M.D. (1977). The effects of frequent testing. Journal of Research and Development in Education, 10, 40-50.

Pedulla, J.J., Airasian, P.W., & Madaus, G.F. (1980). Do teacher ratings and standardized test results of students yield the same information? American Educational Research Journal, 17, 303-307.

Pedulla, J.J., Airasian, P., Madaus, G., & Kellaghan, T. (1977). Proportion and direction of teacher rating changes of pupils' progress attributable to standardized test information. Journal of Educational Psychology, 69, 702-709.

Pestalozzi, H. Uber den Aufenthalt in Stans (1807). In L.W. Seyffarth (ed.) Pestalozzi's samtliche Werke, Bd. 8. Liegnitz: Seyffarth, 1900.

Poole, R.L. (1979). Evaluating and victimizing elementary school children. Education, 97, 115-120.

Poole, R.L. (1976). A teacher-pupil dilemma: Student evaluation and victimization. Adolescence, 11, 341-347.

Popham, W.J. (1973). Establishing performance standards. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Purkey, S.C., & Smith, M.S. 1983). Effective schools: A review. Elementary School Journal, 83, 427-452.

Remmers, H.H., Gage, N.L., & Rummel, J.F. (1960). A practical introduction to measurement and evaluation. New York: Harper and Brothers.

Rheinberg, F. (1983). Achievement evaluation: A fundamental difference and its motivational consequences. Studies in Educational Evaluation, 9, 185-194.

Rimmington, G.T. (1977). Evaluation in history and the social sciences: the longitudinal aspect and its problems. History and Social Science Teacher, 12, 207-211.

Roeder, H.H. (1973). Teacher education curriculum--your final grade is F. Journal of Educational Measurement, 10, 141-143.

Rosenholtz, S.J., & Rosenholtz, S.H. (1981). Classroom organization and the perception of ability. Sociology of Education, 54, 132-140.

Rosenholtz, S.J., & Wilson, B. (1980). The effect of classroom structure on shared perceptions of ability. American Educational Research Journal, 17, 75-82.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. Pp. 122-183 in R.M.W. Travers (ed.), Second handbook of research on testing. Chicago: Rand McNally. 122-183.

Rosswork, S.G. (1977). Goal setting: The effects on an academic task with varying magnitudes of incentive. Journal of Educational Psychology, 69, 710-715.

Roth, S., & Bootzin, R.R. (1974). The effects of experimentally induced expectancies of external control: An investigation of learned helplessness. Journal of Personality and Social Psychology, 28, 253-264.

Rudman, H.C., Kelly, J.L., Wanous, D.S., Mehrens, W.A., Clark, C.M., & Porter, A.C. (1980). Integrating assessment with instruction: A review (1922-1980). East Lansing, MI: Institute for Research on Teaching, Michigan State University, College of Education.

Rudman, M.K. (1978). Evaluating students: How to do it right. Learning, 7, 50-53.

Ryan, K.M., & Levine, J.M. (1981). Impact of academic performance pattern on assigned grade and predicted performance. Journal of Educational Psychology, 73, 386-392.

Salganik, L.H. (1982). The effects of effort marks on report card grades. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Sartore, R.L. (1975). Grading: A searching look. Educational Leadership, 32, 261-275.

Schunk, D.H. (1983). Reward contingencies and the development of children's skills and self-efficacy. Journal of Educational Psychology, 75, 511-518.

Scott, W.R. (1981). Organizations: Rational, natural and open systems. Englewood Cliffs, NJ: Prentice Hall.

Scriven, M. (1978). How to anchor standards. Journal of Educational Measurement, 15, 273-275.

Seligman, M.E.P. (1975). Helplessness: On depression, development, and death. San Francisco: Freeman, 1975.

Shephard, L.A. (1976). Setting standards and living with them. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Simpson, C. (1981). Class room structure and the organization of ability. Sociology of Education, 54, 120-132.

Slavin, R.E. (1977). Classroom reward structure: An analytical and practical review. Review of Educational Research, 47, 633-650.

Slavin, R.E. (1978). Separating incentives, feedback, and evaluation: Toward a more effective classroom system. Educational Psychologist, 13, 97-100.

Smith, L.R. (1984). Effect of teacher vagueness and use of lecture notes on student performance. Journal of Educational Research, 78, 68-74.

Solo, L. (1977) What we do because testing doesn't work. National Elementary Principal, 56, 63-66.

Sorotzkin, F., Fleming, E., & Anttonen, R. (1974). Teacher knowledge standardized test information and its effects on pupil I.Q. and achievement. Journal of Experimental Education, 43, 79-85.

Starch, D., & Elliott, E.C. (1912). Reliability of the grading of high school work in English. The School Review, 20, 442-457.

Stewart, W.J. (1975). A multi-dimensional evaluating-reporting system in the elementary school. Reading Improvement, 12, 174-176.

Stewart, L.G., & White, M.A. (1976). Teacher comments, letter grades and student performance: What do we really know? Journal of Educational Psychology, 68, 488-500.

Stockhard, J., Lang, D., & Wood, W. (1985). Academic merit, status variables, and students' grades. Journal of Research and Development in Education, 18, 12-20.

Symonds, P.M. (1925). Notes on rating. Journal of Applied Psychology, 7, 188-195.

Terwilliger, J.G. 1977). Assigning grades -- philosophical issues and practical recommendations. Journal of Research and Development in Education, 10, 21-39.

Thompson, J.D. (1967). Organizations in action. New York: McGraw-Hill.

Thorndike, R.L. (1969). Marks and marking systems. Pp. 759-766 in R.L. Ebel (ed.), Encyclopedia of educational research. New York: Macmillan.

Thornton, J.W., & Jacobs, P.D. (1972). The facilitating effects of prior inescapable unavoidable stress on intellectual performance. Psychometric Science, 26, 265-271.

Tolor, A., Scarpetti, W.L., & Lane, P.A. (1967). Teachers' attitudes toward children's behavior revisited. Journal of Educational Psychology, 58, 175-180.

Tyler, R.W. (1973). Testing for accountability. In A.C. Ornstein (ed.), Accountability for teachers and school administrators. Belmont, CA: Feardon.

Varenne, H., & Kelly, M. (1976). Friendship and fairness: Ideological tensions in an American high school. Teachers College Record, 77, 601-614.

Waller, W. (1932). The sociology of teaching. New York: Wiley.

Walling, D.R. (1975). Designing a "report card" that communicates. Educational Leadership, 32, 258-260.

Ward, J.G. (1981). Testing and teaching: Partners in learning. Peabody Journal of Education, 58, 91-95.

Warries, E. (1982). Relative measurement and the selective philosophy in education. Evaluation in Education, 5, 191-202.

Weiner, B. (1979). A theory of motivation for some classroom experiences. Journal of Educational Psychology, 71, 3-25.

Webster, M., & Entwisle, D.R. (1979). Expectation effects on performance expectations. Social Forces, 55, 493-503.

Williams, R.G., Pollack, M.J., & Ferguson, N.A. (1975). Differential effects of two grading systems on student performance. Journal of Educational Psychology, 67, 253-258.

Wilson, R.J. (1977). Three faces of evaluation: Students, teachers, curriculum. History and Social Science Teacher, 12, 203-206.

Wilson, S. (1976). You can talk to teachers: Student-teacher relations in an alternative high school. Teachers College Record, 78, 77-100.

Wise, R.I., & Newman, B. (1975). The responsibilities of grading. Educational Leadership, 32, 253-256.

Wortman, C.B., & Brehm, J.W. (1975). Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. In L. Berkowitz (ed.), Advances in Experimental Social Psychology, (Volume 8). New York: Academic Press.

Yeh, J.P. (1978). Test use in schools. Washington, DC: U.S. Department of Health, Education and Welfare and National Institute of Education.

Zahorik, J.A. (1968). Classroom feedback behavior of teachers. Journal of Educational Research, 62, 147-150.

Figure 1

A Model of Evaluation Processes in Schools and Classrooms

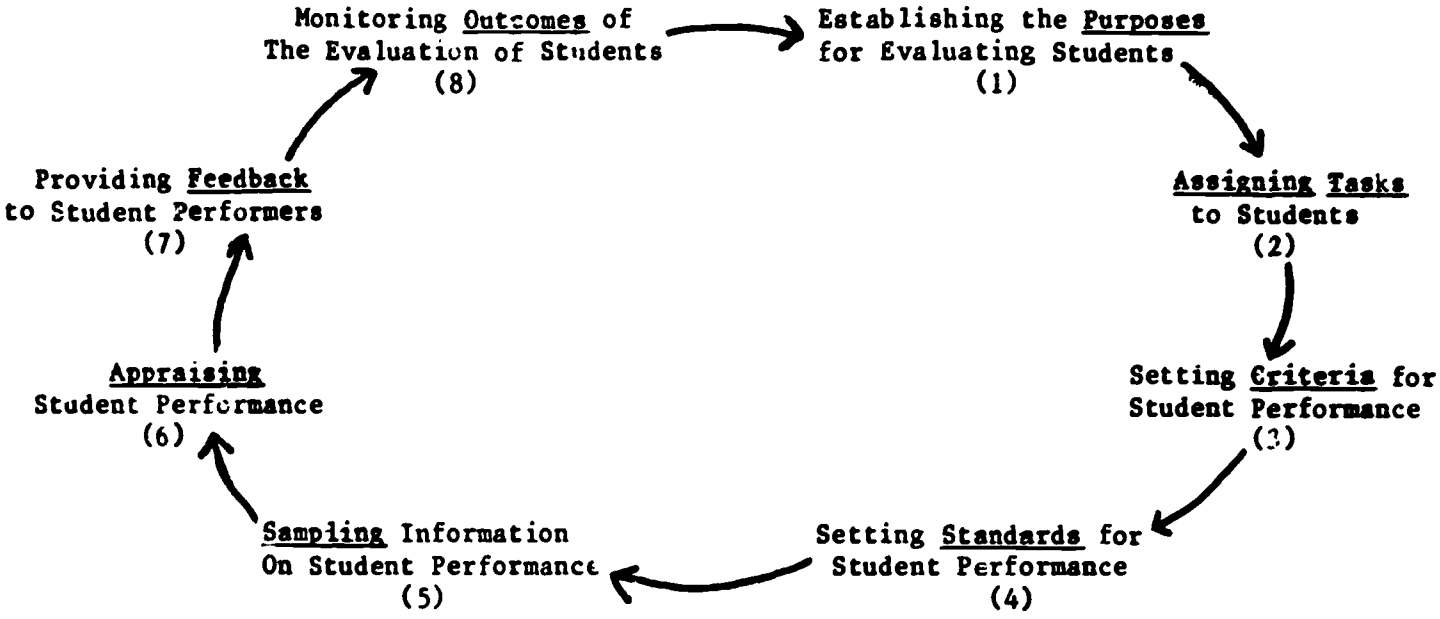

Table 1

List of the Purposes of Evaluation in Schools and Classrooms

1) to assess educational equity (Airasian and Madaus, 1983)

2) to produce evidence on school and program effectiveness, curricular methods, procedures, etc. (Airasian & Madaus, 1983; Ward, 1981; Ahmann and Glock, 1967; Lien, 1967)

3) to guide funds allocation (Airasian and Madaus, 1983)

4) to evaluate teachers (Airasian & Madaus, 1983)

5) to classify students to assign them to particular programs, and provide instructional guidance (Airasian and Madaus, 1983; Fennessey, 1973; Lien, 1967; Remmers, Gage, and Rummel, 1960)

6) to assess competencies to certify successful high school completion and grade promotion (Hambleton, Swaminathan, Algina, & Coulson, 1978; Jackson, 1975: Johnson, 1984; Levine, 1976)

7) to structure better teaching procedures and improve instruction (Fuchs, Deno, and Mirkin, 1984; Ward, 1981; Linn, 1983)

8) to provide better feedback to students and allow them to discover their own abilities, their strengths and weaknesses (Fuchs, Deno, & Mirkin, 1984; Wilson, 1977, Linn, 1983)

9) to monitor individual progress (Hambleton, Swaminathan, Algina, & Coulson, 1978)

10) to diagnose learning deficiencies (Hambleton, Swaminathan, Algina & Coulson, 1978; Ahmann & Glock, 1967)

11) to select students for certain educational and occupational opportunities (Levine, 1976; Lien, 1967)

12) to motivate students (Wise and Newman, 1975: Lien, 1967)

13) to report to parents (Wise and Newman, 1975)

14) to provide feedback to teachers on what students have and have not learned; to guide future teaching (Linn, 1983; Lien, 1967)

15) to flag or identify those items in a curriculum that are particularly important (Linn, 1983)

16) to establish standards and maintain standards (Sartore, 1975; Lien, 1967)

17) to select students for limited positions in programs, institutions, and occupations (Ward, 1981; Warries, 1982; Levine, 1976)

18) to predict future academic success (Wilson, 1977; Warries, 1982)

19) to assess the academic achievement of individual pupils (Ahmann & Glock, 1967)

20) to assess the educational progress of large populations to guide educational policy (Ahmann and Glock, 1967)

21) to furnish instruction to students (Lien, 1967)

22) to adapt instruction to the different needs of individual students (Remmers, Gage, and Rummel, 1960)

23) to provide personal (educational, vocational, social, emotional guidance to students (Remmers, Gage, and Rummel, 1960)

24) to improve public relations through reports to parents and staff (Remmers, Gage, & Rummel, 1960)

25) to enforce the authority and control of the school over students (Natriello, 1982)

Table 2

Percentages of Principals Reporting The Use of Test Results and Other Information on Student Performance as Crucial or Important for Specific Purposes in the School By School Level (Elementary/Secondary) (As Reported by Herman and Dorr-Bremme, 1984)

Purpose	Tests and Other Information Sources					
	Norm-Referenced Tests	Minimum Competency Tests	District Objectives-Based Tests	Teachers' Tests and Assignments	Teachers' Opinions/ Judgements	Other Sources
Curriculum Planning	78/74	60/75	65/57	72/63	88/84	--/--
Assigning Students to Classes	47/72	30/64	38/45	74/75	84/80	49[a]/76[f]
Teacher Evaluation	16/20	11/15	25/21	40/43	--/--	100[b]/95[b]
Allocating Funds	28/24	21/28	29/21	--/--	81/94	77[c]84[c]
Student Promotion	51/24	36/48	48/26	84/84	96/76	94[d]/96[f]
Informing the Public	72/74	38/63	41/43	42/47	--/--	--/--
Communicating to Parents	78/79	56/69	63/45	98/96	95/94	92[e]/97[f]
Reporting to District	81/86	55/72	58/56	53/60	--/--	--/--

--- not asked
a = students' past classroom behavior
b = observations of teachers' teaching
c = specific directions from district
d = classwork throughout the year
e = observations of the student
f = student's report card grades

Table 3

Number and Content of Rating Categories in Reports Used by 312 School Districts Surveyed by Chansky (1975)
(Reported in Percent by Grade Level for Academic and Dispositional Categories)

	Grade Levels				
	K	1-3	4-6	7-9	10-12
Categories	Acad/Disp	Acad/Disp	Acad/Disp	Acad/Disp	Acad/Disp
A) of Steps					
1	13/11	--/--	--/43	--/20	--/24
2	10/--	14/35	--/12	12/20	10/14
3	44/19	33/27	19/26	11/40	--/2[illegible]
4	--/--	15/15	10/9	--/--	--/--
5	--/--	31/20	60/9	69/13	79/24
B) Content					
Adequacy	32/16	42/67	56/30	20/50	11/37
Position	--/--	--/--	16/60	16/15	19/18
Prestige	--/--	--/--	12/--	28/23	23/23
Passage	--/--	--/--	--/--	26/--	31/10
Presence	16/--	--/--	--/--	--/--	--/--
Endorsement	--/--	--/--	--/--	--/--	16/12

Adequacy (e.g., satisfactory)
Position (e.g., average, above average, below average)
Prestige (e.g., excellent, outstanding)
Passage (e.g., pass-fail)
Presence (e.g., all of the time, frequently, not yet)
Endorsement (e.g., superior, good, poor, inferior)

Table 4

The Implications of Selected Programs and Policies on Aspects of the Evaluation Prccess in Schools and Classrooms

Programs & Policies	Elements of the Evaluation Process						
	Purposes	Tasks	Criteria	Standards	Samples	Appraisal	Feedback
Minimum Competency Testing	Certification	Simple	Time Bound	Absolute	Infrequent	Removed from Teacher	Simple Pass/Fail
Mastery Learning	Direction	Small	Not Time Bound	Absolute	Frequent	A's for Mastery	Differentiat Frequent
P.L. 94-142	Direction	Individual-ized	Not specified	Individually Referenced	Frequent	Teacher Dependent	Frequent